Mathematically Optimized, Recursive Prepartitioning Strategies for k-Anonymous Microaggregation of Large-Scale Datasets

Esteve Pallarès[a], David Rebollo-Monedero[a], Ana Rodríguez-Hoyos[a,b], José Estrada-Jiménez[a,b,*], Ahmad Mohamad Mezher[c], Jordi Forné[a]

[a] *Department of Telematic Engineering, Universitat Politècnica de Catalunya (UPC), E-08034 Barcelona, Spain*
[b] *Departamento de Electrónica, Telecomunicaciones y Redes de Información, Escuela Politécnica Nacional (EPN), Ladrón de Guevara, E11-253 Quito, Ecuador*
[c] *Department of Electrical and Computer Engineering, University of New Brunswick, Fredericton, Canada*

Abstract

The technical contents of this work fall within the statistical disclosure control (SDC) field, which concerns the postprocessing of the demographic portion of the statistical results of surveys containing sensitive personal information, in order to effectively safeguard the anonymity of the participating respondents. A widely known technique to solve the problem of protecting the privacy of the respondents involved beyond the mere suppression of their identifiers is the k-anonymous microaggregation. Unfortunately, most microaggregation algorithms that produce competitively low levels of distortions exhibit a superlinear running time, typically scaling with the square of the number of records in the dataset.

This work proposes and analyzes an optimized prepartitioning strategy to reduce significantly the running time for the k-anonymous microaggregation algorithm operating on large datasets, with mild loss in data utility with respect to that of MDAV, the underlying method. The optimization strategy is based on prepartitioning a dataset recursively until the desired k-anonymity parameter is achieved. Traditional microaggregation algorithms have quadratic computational complexity in the form $\Theta(n^2)$. By using the proposed method and fixing the number of recurrent prepartitions we obtain subquadratic complexity in the form $\Theta\left(n^{3/2}\right)$, $\Theta\left(n^{4/3}\right)$,..., depending on the number of prepartitions. Alternatively, fixing the ratio between the size of the microcell and the macrocell on each prepartition, quasilinear complexity in the form $\Theta(n \log n)$ is achieved. Our method is readily applicable to large-scale datasets with numerical demographic attributes.

Keywords: data privacy, statistical disclosure control, k-anonymity, microaggregation, optimized prepartitioning, large-scale datasets

1. Introduction

Thanks to modern information and communication technologies, vast quantities of detailed information, often referred to as big data, are made available to ever more sophisticated and powerful information systems, in order to achieve an unprecedented level of intelligent behavior and personalization. In a wide variety of fields, more utility can be mined from data to unveil qualitatively superior insight into challenges and opportunities that may otherwise remain undiscovered (Halevy et al. (2009); Rosnow & Rosenthal (1989)). For instance, the combination of automatic learning algorithms and the increasing availability of data is leading to remarkable scientific feats such as a better cancer detection.

*Corresponding author.
Email addresses: esteve@entel.upc.edu (Esteve Pallarès), david.rebollo@entel.upc.edu (David Rebollo-Monedero), ana.rodriguez@epn.edu.ec (Ana Rodríguez-Hoyos), jose.estrada@epn.edu.ec (José Estrada-Jiménez), ahmad.mezher@unb.ca (Ahmad Mohamad Mezher), jforne@entel.upc.edu (Jordi Forné)

Nowadays, machine-learning algorithms are being developed to automatically discover such useful "anomalies" in medicine, but they still require vast amounts of data to achieve actionable accuracy. Combining such technologies with big data may lead to truly remarkable scientific feats such as a better cancer detection (Wang et al. (2016); Cukier (2014)). In fact, human proficiency is being combined with machine-based mechanisms to provide augmented intelligence from large-scale databases.

But the revolutionary advances accomplished in the big data era poses equally serious privacy risks. Although identifiers are typically suppressed from shared or published data, there remain the so-called *quasi-identifier* attributes, essentially publicly available demographic attributes which, when combined, can be used to re-identify individuals (Sweeney (2000a); Narayanan & Shmatikov (2008); AOL). In fact, it was shown by Sweeney (2000a) that 87% of the population in the United States could be unequivocally identified solely on the basis of the triple consisting of their date of birth, gender and 5-digit ZIP code, according to 1990 census data This re-identification might enable privacy attackers to link the identity of subjects with their corresponding sensitive attributes.

To reduce this disclosure risk in microdata files (individual user data tabulated in records), *statistical disclosure control* (SDC) is commonly used. Accordingly, SDC mechanisms build on perturbing quasi-identifier attributes to de-identify records; a process also called anonymization. The privacy models enforced through user data perturbation, e.g., k-anonymity (Sweeney (2000a); Samarati (2001)) or ϵ-differential privacy (Dwork (2006)), are usually conditioned by a privacy parameter that defines an upper bound on the re-identification risk. However, in practice, other parameters such as data utility and mechanism usability convolute the task of protecting privacy. Evidently, data perturbation comes at the cost of some loss in data utility. Additionally, finding a balance between privacy and utility, when big data is involved, might turn private data analysis unfeasible or unusable for some applications where, e.g., mechanisms must execute in a reasonable amount of time despite the size of the data.

Differential privacy and other privacy criteria such as multi-party computation (Vaidya et al. (2006); Dankar et al. (2014)) and integral privacy (Torra & Navarro-Arribas (2018)) are out of the scope of this work, since our target application is that of data release for general statistical analysis with a focus on data utility. Recall that differential privacy is conceived for online querying on predefined computations, and that in general it imposes stringent restrictions, both in terms of usability and data utility. Those restrictions, introductorily explained also by Matwin et al. (2015), render it rather inadequate for our purposes.

k-Anonymous microaggregation is a high-utility mechanism to protect privacy in microdata by obfuscating demographic attributes. Carefully aggregating these attributes, a minimum level of distortion must be applied to original data. In fact, k-anonymous microaggregation is an excellent approach to applications requiring the preservation of data utility (Rodríguez-Hoyos et al. (2018)). Unfortunately, current microaggregation algorithms entail a very high computational cost when anonymizing big data, which derives in longer significantly running times.

In this work, we propose and analyze a strategy to significantly reduce the running time of the k-anonymous microaggregation algorithm when applied on large datasets. By recursively prepartitioning a dataset, under certain conditions, our method can turn the quadratic computational complexity of microaggregation into quasilinear complexity, while the resulting distortion increases moderately, at least for a few prepartitioning steps. Said otherwise, we optimize a prepartitioning strategy of data to significantly reduce the running time of microaggregation algorithms, having a relatively small impact on the resulting utility of data.

In brief, the fundamental philosophy behind this paper is to model computational demands and decide the best, mathematically optimized, course of action. Evidently, an optimal decision pays off under circumstances where the risk of a suboptimal or even naïve approach is highest. We believe that, for instance, in the context of big data and real-time applications, such risk is very high, in particular due to the quadratic complexity of popular microaggregation algorithms.

1.1. Fundamentals of Statistical Disclosure Control and Microaggregation

SDC concentrates on protecting the privacy of individuals or organizations whose information is originally presented as a microdata set (a database table whose records carry information concerning identifiable

subjects). Such representation implies an individual record associated with each data subject; each record contains a set of attributes of three different types: identifiers, quasi-identifiers, and confidential.

- *Identifiers* unequivocally identify respondents in the microdata set. Examples of identifiers are full names or medical record numbers. They are commonly removed before publishing the microdata set, in order to guarantee anonymity.
- *Quasi-identifiers* may include demographic attributes such as age, gender, address or physical attributes, which combined or linked with other external information can be used to re-identify data subjects.
- *Confidential attributes* contain sensitive information on the respondents, such as salary, political affiliation, and health condition.

In practice, the strong identifying capabilities of a few demographic (quasi-identifier) attributes have rendered the mere suppression of identifiers grossly insufficient to effectively protect the privacy of data subjects (Sweeney (2000b,a)). Thus, further perturbation of data is required to effectively guarantee their anonymity.

In Fig. 1, we illustrate how a perturbed, and thus more private, version of a dataset is obtained to be published instead of the original one. The original dataset combines attributes common in medical surveys. Four quasi-identifiers are shown, sex, age, weight, and body mass index, and four confidential attributes, heart rate, blood pressure, oxygen saturation, and electrocardiogram. The figure at hand shows how, in order to preserve the privacy of respondents, perturbation is applied to quasi-identifiers.

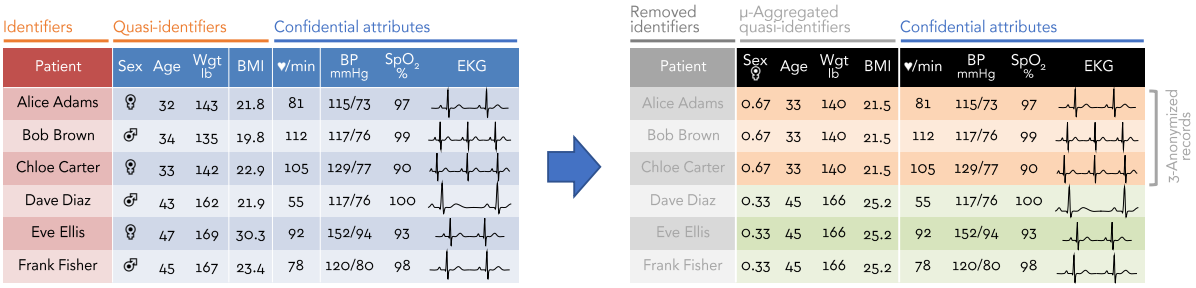


Figure 1: Example of k-anonymous microaggregation of published data with $k = 3$. Identifiers are first suppressed. The quasi-identifiers (demographic attributes) in the left table are anonymized on the right. Confidential attributes (heart rate, blood pressure, oxygen saturation, and electrocardiogram) are left intact after anonymization since the rest of the data is de-identified. By grouping quasi-identifiers, individuals remain demographically indistinguishable among a group of $k = 3$ uncertain possibilities.

This technique is called microaggregation, which is applied to enforce k-anonymity. k-Anonymity is a privacy model that guarantees that each tuple of quasi-identifiers is identically shared by at least k records in a dataset (Sweeney (2002)). Rather than making the original table available, a perturbed version is published where aggregated records of quasi-identifying values are replaced by a common representative tuple. The result is a microaggregated dataset that may prevent re-identification attacks. The *de facto* standard for numerical microaggregation is the maximum distance to average vector algorithm (MDAV). It was proposed by Domingo-Ferrer & Torra (2008) as a practical evolution of a multivariate fixed-size microaggregation method and conceived by Sankar et al. (2013).

1.2. Contribution and Organization

Although traditional microaggregation provides an efficient method to anonymize while introducing reasonably low amounts of distortion, it proves expensive in terms of computation time when it is applied on large-scale datasets. The leading objective of this contribution is to provide a faster k-anonymization method to address this issue. To do so, we apply an optimized prepartitioning method using the microaggregation algorithm, recursively, thus, creating in each iteration clusters that have fewer records than the previous one. In the last iteration, the desired k-anonymity is achieved obtaining clusters with at least k records each. The running time of the entire process can be optimized by choosing the number of records

per cluster in each iteration, and the number of iterations appropriately. Albeit our work is illustrated with the special case of the widely used algorithm known as maximum distance to average vector (MDAV), the method outlined would be readily applicable to other microaggregation techniques.

The proposed method is able to microaggregate databases significantly faster than MDAV, mainly for large datasets, with a little impact on information utility. However, there is a tradeoff between the reduction of the running time and the loss of data utility. Based on our experimental results, we propose a range of values for both, the number of iterations, and the number of registers per cluster in each iteration. The aim is to reduce the loss of information utility in exchange for increasing the optimal running time. The highlights of our contribution are summarized in Fig. 2.

More concretely, our contributions are the following.

- We propose a novel method to microaggregate large datasets, by using a mathematically optimized prepartitioning strategy tailored to certain microaggregation algorithms with superlinear complexity.
- The strategy is applied recursively, until the desired k-anonimity is achieved, which considerably reduces the running time compared to the conventional algorithm.
- Analytical expressions for the number of records per cluster and the number of recursions that optimize the running time have been obtained.
- Using the optimal number of records per cluster and fixing the number of recursions, the quadratic computational complexity $\Theta(n^2)$ of conventional microaggregation is gradually reduced to subquadratic complexities of $\Theta\left(n^{3/2}\right)$ for one recursion, $\Theta\left(n^{4/3}\right)$ for two recursions, and so on. In general, the complexity $\Theta(n^{(j+1)/j})$ attained decreases with the number $j \geqslant 1$ of recursive iterations, approaching the unit exponent in the limit. Each reduction step represents a substantial speed-up in the anonymization process for a large number n of records.
- Alternatively, fixing the ratio between the size of the microcell and the macrocell on each prepartition, quasi-linear computational complexity $\Theta(n \log n)$ is achieved, drastically improving the quadratic complexity $\Theta(n^2)$ of microaggregation without prepartitioning.
- As the number of records per cluster that optimize the running time in each iteration is not an integer value, an expression for the relative error of the optimal running time has been also computed.
- In order to validate the theoretical results of our proposal, a synthetic and a real dataset have been microaggregated with our novel optimized prepartitioning strategy. In both cases our method dramatically reduces the running time, closely matching the predictions of our theoretical models.
- We have analyzed the impact of our proposal on the utility of the data of microaggregated, specifically, in terms of squared distortion. The experimental results show that for a small number of subsequent prepartitions, our method offers substantial time gains at the expense of negligible distortion degradation. However, aggressive approaches with high number of iterations may lead to considerable distortion impact.
- Additionally, the effect of the chosen number of records per partitioning cluster on data utility has been analyzed. Experimentally, we have observed that the values of practical interest in terms of both running time and distortion are moderately higher than those corresponding to the theoretical optimization. More precisely, we investigate the trade-off between relative time gain τ and relative distortion degradation δ with respect to conventional microaggregation. We show the lower envelope on the τ-δ plane, for various modes of parametric operation of our prepartitioning method.

The rest of this paper is organized as follows. §2 briefly reviews the current state of the art in k-anonymous microaggregation metrics and algorithms in the SDC literature. Some related works are presented in §2. §3 formally presents the proposed formulation of the optimized prepartitioning algorithm, while §4 presents the experimental analysis and outcomes of the proposed algorithm. Finally, conclusions are drawn in §5.

2. State of the Art on k-Anonymous Microaggregation

Microaggregation is a mechanism that aims at protecting the privacy of individuals whose personal data is released in a microdata set. To do so, the quasi-identifier attributes are perturbed in such a way that

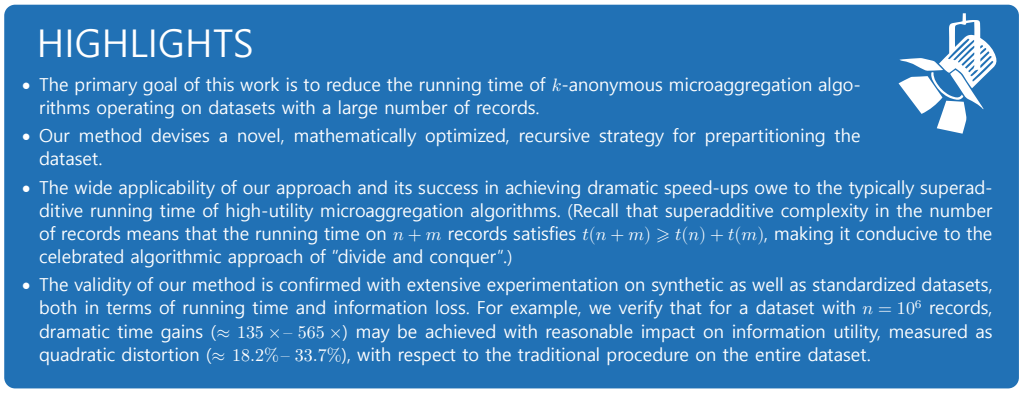


Figure 2: Highlights of our contribution.

k-anonymity (Samarati (2001); Sweeney (2000b)) is satisfied. This privacy model guarantees that each individual's information contained in a released dataset cannot be distinguished from that of at least $k-1$ individuals whose information also appears in the dataset. Microaggregation was adopted as a k-anonymous-based mechanism in (Defays & Nanopoulos (1993); Domingo-Ferrer & Mateo-Sanz (2002); Domingo-Ferrer et al. (2008); Domingo-Ferrer & Torra (2005)).

If tuples of quasi-identifier attributes in a dataset could be represented as points in the Euclidean space, k-anonymous microaggregation would consist in partitioning these points in cells of size k, and quantizing each cell and its elements with a representative point. Perturbed quasi-identifiers would be characterized by the set of representative points. This is graphically illustrated in Fig. 3.

2.1. Shortcomings of k-Anonymity as Privacy Criterion

Although the criterion of k-anonymity is very popular, it has a weakness: it operates only in quasi-identifiers so neglects by default all the rest of information available for an attacker. Namely, k-anonymity does not consider, e.g., the statistical properties of confidential attributes (and thus their disclosure potential), both in the dataset and in the entire population. Since this attacker's knowledge is overlooked, similarity, skewness or background-knowledge attacks become feasible (Domingo-Ferrer & Torra (2008); Rebollo-Monedero et al. (2010, 2013b)).

Facing these issues, additional privacy criteria have been proposed in the literature. For instance, trying to tackle skewness and similarity attacks, p-sensitive requires that each group of k-anonymized records contains at least p different values of each confidential attribute (Truta & Vinay (2006); Sun et al. (2008)). In a broader approach, l-diversity proposes that each group have at least l well-represented confidential values. Unfortunately, none of these criteria guarantees complete protection if confidential attributes within a k-anonymous group are semantically similar.

Similarity and skewness attacks could also arise if the distribution of confidential attributes within a k-anonymous group differs from that within the original dataset. Accordingly, other privacy criteria such as t-closeness by Li et al. (2007), delta-disclosure by Brickell & Shmatikov (2008), and average privacy risk (Rebollo-Monedero et al. (2008, 2010)) pose additional requirements in the distribution of confidential attributes within groups. The aim is that confidential attributes in each group of anonymized records are stratified according to their distribution in the original dataset.

2.2. Algorithms for k-Anonymous Microaggregation

Getting groups of exactly k records from a microdataset is a strong restriction to satisfy k-anonymity. In fact, multivariate microaggregation is an NP-hard problem. Thus, several heuristic algorithms have been proposed to cope with such complexity. First, the maximum distance (MD) (Domingo-Ferrer et al. (2009)) and its variation, maximum distance to average vector (MDAV) (Domingo-Ferrer et al. (2009); Domingo-Ferrer & Torra (2005)) are catalogued as fixed-size algorithms because all aggregated groups but one have

exactly k elements. Variable-size algorithms include, on the other hand, the μ-Approx by Gursoy et al. (2017), the minimum spanning tree (MST) by Hundepool et al. (2003), the variable MDAV (V-MDAV) by Inan et al. (2009) and the two-fixed reference points algorithms (TFRP).

The *de facto* standard for numerical microaggregation is the MDAV algorithm. It was proposed by Hundepool et al. (2003) as a practical evolution of a multivariate fixed-size microaggregation method and conceived by Domingo-Ferrer & Mateo-Sanz (2002). Since we use MDAV to illustrate our novel methods in this work, for the sake of reproducibility, we provide in Algorithm A a simplified version of that given by Domingo-Ferrer & Torra (2005) and termed "MDAV generic".

In general, the implementations of microaggregation have been oriented to preserve the utility of data (Lin et al. (2010); Matatov et al. (2010); Domingo-Ferrer & González-Nicolás (2010)), which is evidently affected due to perturbation. Although the usual metric to measure such utility is SSE, other semantic-oriented metrics could be considered, aiming to conceive realistic implementations. For instance, Li et al. (2007) explore an elegant extension of the usual SSE metric that contemplates not only the distortion of the quasi-identifiers due to aggregation, but also the valuable statistical dependence between quasi-identifiers and confidential attributes, in order to improve the statistical reliability of demographic studies. Furthermore, in this line of involving the inherent relationship between quasi-identifier and confidential attributes, a machine learning based metric is used by Rodríguez-Hoyos et al. (2018) to systematically determine the impact (surprisingly limited) of microaggregation on the practical utility of data.

We use MDAV since it is a well-known microaggregation algorithm for numerical data in the literature of database anonymization. In fact, many of these works refer to MDAV not only as a standard method (or the most widely used) for microaggregation (Templ (2017); Mahmood et al. (2012)) and use it as a baseline for comparison purposes (Sun et al. (2012); Mortazavi & Jalili (2017)), but also recommend it due to its efficiency and performance (Templ et al. (2014)) in terms of the resulting data utility. Even in recent years, MDAV is used as the baseline to find new and improved microaggregation approaches (Iftikhar et al. (2019); Salas & Torra (2018); Liu et al. (2018); Fayyoumi & Nofal (2018); Wei et al. (2018); Zhang et al. (2018)).

Talking about its impact on data utility, MDAV is even being actively used to enhance the utility of differentially private data sets via record masking (Parra-Arnau et al. (2019); Sánchez et al. (2016); Soria-Comas et al. (2014)). Interestingly, its averaging operations to find a representative centroid turn to be a mechanism to reduce the amount of noise required to meet a differential privacy criteria.

Beyond the performance of MDAV in terms of data distortion, the key of our proposal lies in the mathematical model and optimization of the parameters employed in prepartitioning of data to obtain, in practice, substantial improvements in running time of this standard solution. We apply the "divide and conquer" principle to reduce the quadratic (n^2) complexity of MDAV to subquadratic ($n^{3/2}$) complexity. Thus, this improvement is even better the greater is the number of records, making our work applicable to big data.

Since k-anonymous microaggregation generally implies partitioning a dataset in groups of size at least k as with MDAV, our novel method could be easily implemented on other microaggregation techniques to reduce their execution times. We are aware that MDAV does not always yield the lowest possible distortion, and we are naturally intrigued by possible qualitative differences in utility degradation when applying our optimized prepartitioning principles to microaggregation algorithms capable of preserving data utility even better than MDAV. Of particular interest are the algorithms tested in Aloise & Araújo (2015). Another candidate for this optimization could be PCL, by Rebollo-Monedero et al. (2013a), inspired by the Lloyd algorithm, which outperforms MDAV in terms of data utility that comes at a cost in running time.

In Abidi & Yahia (2017), we can find a compelling example for the specific case of the standardized dataset "Census" and the common value $k = 10$ of the anonymity parameter, where the MSE distortion produced by traditional MDAV (without any form of prepartitioning) is $\mathcal{D} \approx 0.142$. This distortion may be further reduced to 0.132, according to Aloise & Araújo (2015). We should mention that to the best of our knowledge, the lowest distortion reported for "Census" and $k = 10$ is $\mathcal{D} \approx 0.122$, obtained with the *probability-constrained Lloyd algorithm (PCL)* proposed by Rebollo-Monedero et al. (2011) as an adaptation of necessary conditions for optimal quantization to k-anonymous microaggregation, as explained in Rebollo-Monedero et al. (2013a), although this last method is computationally demanding. Even though the optimality conditions built into the design of PCL are necessary but not sufficient, in practice, PCL offers excellent data utility in numerical

k-anonymous microaggregation, even below that of MDAV. For this particular example, the distortion of 0.122 achieved by PCL is approximately 13.7% smaller than that of 0.142 obtained with MDAV.

2.3. Prepartitioning as a Mechanism for Computational Improvement

Regardless of the metric used, preserving utility commonly derives in more sophisticated and significantly costlier implementations of microaggregation in terms of computational time (Rebollo-Monedero et al. (2011)). Fortunately, microaggregation algorithms have proven to be susceptible to computational improvements (Mohamad Mezher et al. (2017)) due to some of their properties. These properties are exploited, for instance, by Rebollo-Monedero et al. (2018), who propose to consider data availability over time in addition to demographic similarity for substantially faster microaggregation, due not only to superadditivity, but also to mathematically optimized scheduling. Interestingly, one of the strategies applied involve prepartitioning the dataset in a small number of large macrocells and the individual postpartitioning of these macrocells into cells of the intended size k. Rebollo-Monedero et al. (2011) implement and evaluate such prepartitioning and, although additional distortion is derived, it proves to be a practically convenient strategy to gracefully trade-off distortion for running time.

When processing data, the concept of data partitioning is pretty known as a step previous to distributing the computing load of a system. Namely, by dividing data into multiple chunks, independent instances are enabled to process each of such pieces, allowing for a more efficient exploitation of (particularly, computing) resources. Evidently, an optimized prepartitioning strategy, i.e., looking for the most convenient way to divide data, will bring the maximum benefits of this "divide-and-conquer" approach. Optimized prepartitioning has been applied in different domains. For instance, Arres et al. (2015) use data prepartitioning and distribution optimization to increase the efficiency of database relational operations such as indexing, grouping, aggregation and joining. Moreover, Tabik et al. (2016) provides a model to find an optimal data partition for applications found on the bioinformatics domain; their objectives span efficiently balancing workload and deactivating slower devices. Still in line with an improved performance of data processing, parallelization is a natural consequence of data prepartitioning, and once again its optimization defines the level of the resulting efficiency in data processing (Ke et al. (2011)).

While prepartitioning is by no means a novel idea, we propose a mathematically optimal procedure to carefully select the size of the partitions or macrocells. Then, we extend this procedure to consider a recursive prepartitioning strategy for a desired number of recursive stages, with optimal intermediate macrocell sizes. Finally, we go one step further and consider the optimization of the number of stages itself. To the best of our knowledge, previous research has not addressed these avenues, so we present them in this work.

Algorithm 1 MDAV "generic", functionally equivalent to Algorithm 5.1 in Domingo-Ferrer & Torra (2005).

function MDAV
input k, $(x_j)_{j=1}^n$ ▷*Anonymity parameter k, quasi-ID portion $x_1, \ldots, x_n \in \mathbb{R}^m$ of a dataset of n records*
output q ▷*Assignment function from records to microcells $j \mapsto q(j)$*
while $2k$ points or more in the dataset remain to be assigned to microcells **do**
  find the centroid (average) C of those remaining points
  find the furthest point P from the centroid C, and the furthest point Q from P
  select and group the $k-1$ nearest points to P, along with P itself, into a microcell, and do the same with the $k-1$ nearest points to Q
  remove the two microcells just formed from the dataset
if there are k to $2k-1$ points left **then**
  form a microcell with those and finish
else ▷*At most $k-1$ points left, not enough for a new microcell*
  adjoin any remaining points to the last microcell ▷*Typically nearest microcell*

2.4. Related Work

There are several works related to adapting or creating better microaggregation algorithms, not only in terms of runtime, but also in terms of data utility. For instance, Laszlo & Mukherjee (2005) raised the idea of an efficient clustering mechanism capable of reasonably dealing with large data sets while preserving the utility of data through a partitioning method of a modified minimum spanning tree. Sun et al. (2012) also proposed an efficient and effective microaggregation approach that outperforms that of MDAV by relying on the concept of entropy to evaluate the amount of mutual information as a distance among records in microdata. In addition, Mortazavi et al. (2014) introduced fast data-oriented microaggregation (FDM), a method capable of getting multiple protected versions of a large data set (for different values of k) in a single load. FDM also offered a better trade-off between information loss and disclosure risk in comparison with similar algorithms.

Interestingly, the specific methods for recursive partitioning we use in our work are similar in spirit to those applied on other proposals such as Mondrian by LeFevre et al. (2006) and Fuzzy Microaggregation by Domingo-Ferrer & Torra (2003) but different in the way clusters are conceived and built.

Although most of these solutions resort to modifying the microaggregation algorithm itself, other proposals may envision post-processing techniques, i.e., that could be applied after any microaggregation algorithm without changing its clustering phase (Mortazavi & Jalili (2017)).

The main difference of our method with respect to those briefly mentioned above lies in the fact that our approach is entirely focused on formally and mathematically optimizing the prepartitioning parameters we modeled to get, in practice, important gains in runtime. Such gains are due to the reduction of time complexity from quadratic to subquadratic and quasilinear. As a consequence, the resulting improvement in execution time is much better than that obtained by other microaggregation algorithms, and improves even more when big data is involved.

As a final note, the optimized prepartitioning strategies developed in this study could be synergically combined with the functional and computational improvements introduced by the two following works. Rebollo-Monedero et al. (2017) proposes a probabilistic generalization of k-anonymous microaggregation in a very novel context: large-scale demographic surveys, in which respondent participation is uncertain. And, Rebollo-Monedero et al. (2019) get reductions in running time and memory usage with negligible impact in information utility by using the algebraic-statistical technique of principal component analysis (PCA), in order to effective reduce the number of attributes to be processed.

3. Formulation of the Problem of Optimized Recursive Prepartitioning

This section introduces the basic notation employed and the fundamentals of multivariate numerical k-anonymous microaggregation. Subsequently, it describes the optimized prepartitioning strategy for efficient anonymization of large-scale datasets.

3.1. Basic Notation and Fundamentals of Multivariate Numerical k-Anonymous Microaggregation

The traditional k-anonymous microaggregation algorithm partitions a set of quasi-identifiers into cells of at least k samples. As previously described by Rebollo-Monedero et al. (2013a, 2011), we formally present microaggregation as a quantization problem. The scope of our analysis is mildly limited to *numerical data*, meaning that we assume that the quasi-identifiers aggregated are represented by n points $X = x_1, \ldots, x_n$ placed in the Euclidean space $\mathbb{R}^m$ of dimension m. These points are grouped into cells indexed by $q = 1, \ldots, Q \leqslant n/k$. Let x_j define the j^{th} record and $\hat{x}_j$ the mean value or centroid of the samples aggregated in the cell where x_j is assigned. The construction of cells, gathering at least k nearby samples, is represented by the cell-assignment or quantization function $q(j)$. Before releasing the dataset, each quasi-identifier tuple x_j is replaced by its perturbed version $\hat{x}_j$, the mean of the corresponding cell $q = q(j)$. This determines a centroid-assignment or reconstruction function $\hat{x}(q)$. This process is conceptually depicted in Fig. 3.

We mentioned in the introductory review of k-anonymous microaggregation that practical algorithms are designed to perturb quasi-identifiers in a way such that the statistical quality of the published data is guaranteed. Technically speaking, microaggregation is similar to a quantization problem: the algorithms

find a partition of the set of quasi-identifying tuples in cells of k elements and try at the same time to reduce the distortion red when replacing each element in a cell by its representative within this cell. Fig. 3 conceptualizes k-anonymous microaggregation as minimum-distortion vector quantization, with the added restriction that cells be at least of size k. The function $q(j)$ assigns the quasi-identifier tuple x_j to microcell q, which will contain at least $k-1$ other points, and whose value will be replaced by the common reconstruction tuple or centroid $\hat{x}(q)$.

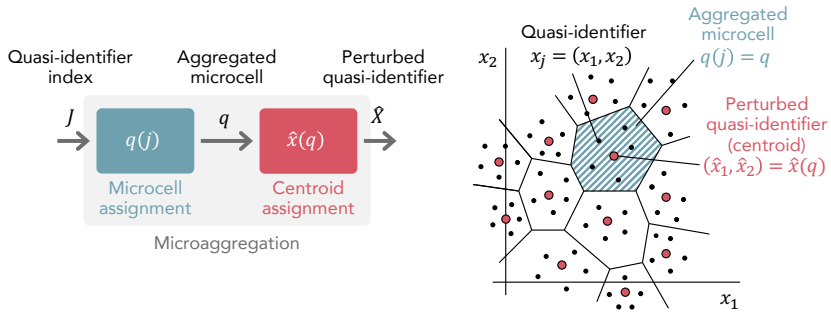


Figure 3: Traditional microaggregation interpreted as a quantization problem on the record indices j, represented by a microcell assignment function $q(j)$, together with a centroid assignment function $\hat{x}(q)$ that reconstructs the perturbed version $\hat{x}_j$ of the original quasi-identifier x_j. The figure also shows an example of microaggregation of 2-dimensional quasi-identifiers with anonymity parameter $k = 5$ (the red point represents the centroid). Each microcell of five points is assigned a representative centroid. In this example, the two-dimensional quasi-identifiers could correspond to a pair of demographic attributes such as age and number of school years. The centroids are the value of the published, perturbed quasi-identifiers within each cell.

Recall the common practice in SDC of normalizing each quasi-identifier. In this work we use *zero-mean unit-variance normalization*. *Unit-variance columnwise normalization* is particularly useful to confer equal importance to each quasi-identifier when microaggregating data. In such a way, we avoid misinterpretation of data, e.g., when arbitrary choices of units are done, say choosing pounds or kilograms for weights, and inches or meters for heights. Of course, reweighting is possible in applications where certain quasi-identifiers are deemed of greater importance than others in the quantification of data utility.

This normalization also implies that the usual measure of *distortion*, precisely, the ratio between the sum of squared errors

$$\mathrm{SSE} \stackrel{\mathrm{def}}{=} \sum_{j=1}^{n} \|x_j - \hat{x}_j\|^2$$

and the sum of squares total

$$\mathrm{SST} \stackrel{\mathrm{def}}{=} \sum_{j=1}^{n} \|x_j\|^2 = mn,$$

matches the usual definition of distortion in the field of vector quantization, as *mean squared error* (MSE) normalized by dimension:

$$\mathcal{D} \stackrel{\mathrm{def}}{=} \frac{\mathrm{SSE}}{\mathrm{SST}} = \frac{1}{mn} \sum_{j=1}^{n} \|x_j - \hat{x}_j\|^2 = \frac{1}{m} \mathrm{E}\, \|X - \hat{X}\|^2.$$

A discussion of the optimality conditions of k-anonymous microaggregation can be found by Rebollo-Monedero et al. (2013a). Let $n(q) \geqslant k$ denote the size of microcell q. While it is well known that the centroid or conditional expectation

$$\hat{x}(q) = \frac{1}{n(q)} \sum_{j \,|\, q(j)=q} x_j = \mathrm{E}\,[X|q]$$

minimizes the MSE within each microcell, and thus it constitutes the optimal reconstruction for a given microcell assignment function $q(j)$, the problem of constructing such microcell assignment, under the restriction that it contains at least k points, may prove difficult. In practice, the k-anonymous microaggregation

algorithm MDAV, introduced in our review of the state of the art in §2, is an excellent heuristic in terms of MSE and is one of the most well-known k-anonymous microaggregation algorithm in the literature.

3.2. Optimized Prepartitioning Strategy Proposal

This subsection formally presents the proposed prepartitioning strategy with the clear objective of reducing the execution time of k-anonymous microaggregation algorithm. Throughout this paper, we have demonstrated that the running time t of the microaggregation algorithm scales with the square of the number n of records and inversely proportional to the anonymity parameter k, according to

$$t = \frac{n^2}{k},$$

where the time units have been chosen to avoid the hassle of a proportionality constant. This is the case for popular microaggregation algorithms such as MDAV by Domingo-Ferrer & Torra (2005). To reduce this time, we propose to partition the dataset recursively in several intermediate steps using the same microaggregation algorithm until the desired k-anonymity is achieved.

3.2.1. A Single-Stage Prepartitioning Case

The single-stage prepartitioning case consists of 2 steps. The first step is prepartitioning the whole dataset into n/k_1 clusters using MDAV followed by a second step of applying MDAV again with anonymity parameter k, on each one of the n/k_1 clusters created in the first step. In this way, all clusters are of almost k records. Obviously, k_1 is an integer between k and n. Now, the total new running time is n^2/k_1 for the first step plus k_1^2/k for each one of the n/k_1 clusters created in the first step. Consequently, the total running time t_2 in this case can be written as

$$t_2 = n\left(\frac{n}{k_1} + \frac{k_1}{k}\right).$$

We use the nomenclature t_2 to indicate a running time in two steps.
For convenience, we normalize the running time t_2 in terms of the number n of records as

$$\frac{t_2}{n} = \frac{n}{k_1} + \frac{k_1}{k}.$$

As we can see, the normalized running time t_2/n is the sum of two functions that depend on k_1 (n and k are considered constants values), a decreasing positive hyperbola function n/k_1 and a positive increasing linear function k_1/k. As a result, we have a convex function with a local minimum where both summands should be equal. Consequently, the optimal value for k_1 that minimizes the running time is

$$k_1^\star = \sqrt{nk}.$$

Note that we use the star symbol as a superscript just to indicate that the variable takes on its optimal value. Finally, the minimum normalized running time using a single-stage prepartitioning is

$$\frac{t_2^\star}{n} = 2\sqrt{\frac{n}{k}}.$$

Probably, $k_1^\star$ will not be an integer value, so in order to be an integer it has to be rounded to the nearest integer value. Now, to calculate the relative error in the running time caused by the rounding issue we write the following

$$\epsilon_t = \frac{t_2(k_1)}{t_2^\star} - 1 = \cosh\ln\frac{k_1}{k_1^\star} - 1.$$

We can clearly observe that the relative error ϵ_t is always a positive value, since $t_2^\star$ is the minimum of all possible values of t_2. Applying the Taylor series expansion on the function $\cosh\ln x$ at $x = 1$, that is, around the optimal value $k_1^\star$, we can approximate the relative error of the running time using the relative error of k_1 as

$$\epsilon_t = \frac{1}{2}\left(\frac{k_1}{k_1^\star} - 1\right)^2 + O\left(\left(\frac{k_1}{k_1^\star} - 1\right)^3\right) \approx \frac{1}{2}\epsilon_k^2.$$

Due to rounding, the maximum relative error ϵ_k is either $1/k_1^\star$ if k_1 is rounded to the immediately higher or lower integer, or $1/(2k_1^\star)$ if it is rounded to the closer integer. In the worst case, the relative error for the running time will satisfy the following inequality

$$\epsilon_t < \frac{1}{2kn}.$$

The exact expression for the relative error can be obtained using $k_1 = k_1^\star(1+\epsilon_k)$ and it is depicted in Fig. 4. Precisely, we have

$$\epsilon_t = \cosh \ln(1+\epsilon_k) - 1 = \frac{\epsilon_k^2}{2(1+\epsilon_k)}.$$

It can be seen that the relative error of the running time is slightly lower if we round k_1 to the upper integer instead of the lower one, however, this relative error due to round k_1 to the upper integer is not far away from the best option which is rounding k_1 to the nearest possible integer.

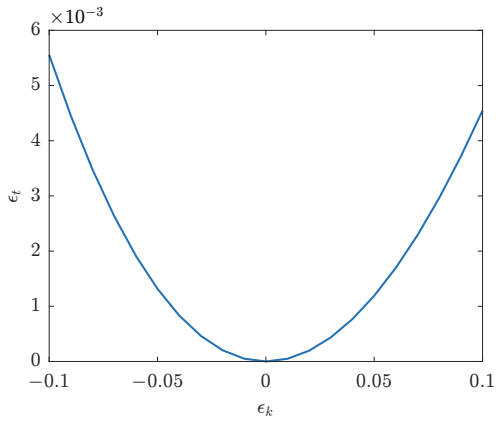


Figure 4: Relative error of the normalized running time as a function of the relative error of k_1, ϵ_k. The error is slightly lower if we round k_1 to the upper integer instead of the lower one, however, this relative error due to round k_1 to the upper integer is not far away from the best option which is rounding k_1 to the nearest possible integer.

In order to compare the running time of our proposal with the conventional microaggregation algorithm, we shall define the relative running time as

$$\tau_2 = \frac{t_2}{t} = \frac{k}{k_1} + \frac{k_1}{n}.$$

One of the advantages of using this parameter τ_2 is that it does not depend on the computer where the algorithm is executed neither the unit of time being used. The two step partitioning algorithm will be better than the conventional one if $\tau_2 < 1$, then the range of good values for k_1 would be $n/2 \pm \sqrt{(n/2)^2 - nk}$ The optimal relative running time (i.e., the minimum one) is

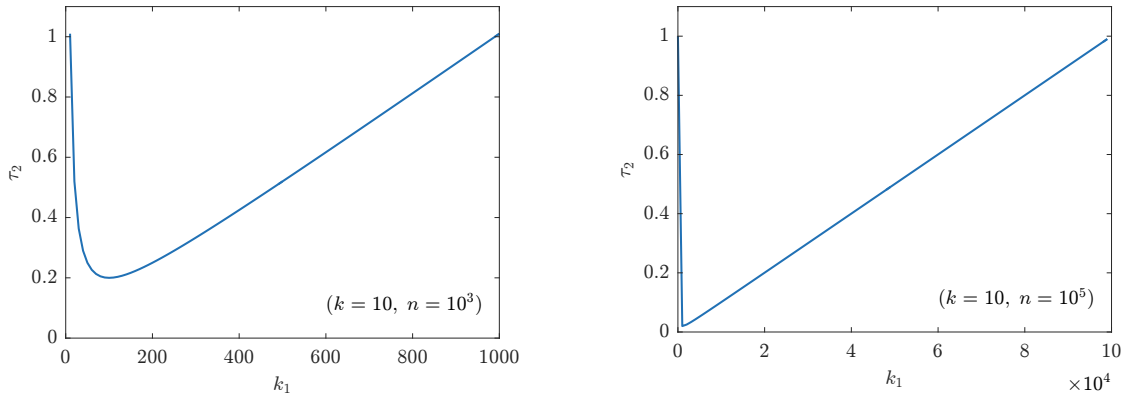
$$\tau_2^\star = 2\sqrt{\frac{k}{n}},$$

which decreases as the number of records of the dataset increases, as shown in Fig. 5. Therefore, our proposal suits better for large datasets rather than small ones.

3.2.2. Multiple-Stage Prepartitioning

The optimization running time can be improved by applying recursively the prepartitioning algorithm for an undetermined number of steps. For instance, the three prepartitioning steps would be achieved by partitioning the dataset into n/k_2 clusters with almost k_2 records, after that, each cluster will be partitioned again into clusters with almost k_1 records, and finally all the clusters will be partitioned into clusters with almost k records. In this case the normalized running time will be

$$\frac{t_3}{n} = \frac{n}{k_2} + \frac{k_2}{k_1} + \frac{k_1}{k}$$



(a) (b)

Figure 5: Theoretical relative running time as a function of k_1, with a single-stage prepartitioning with $k = 10$ for a dataset with (a) 10^3 records, and (b) 10^5 records.

For a generic case with j steps, the normalized running time can be written as

$$\frac{t_j}{n} = \sum_{i=1}^{j} \frac{k_i}{k_{i-1}}.$$

In the above expression, we have assumed that $k_0 = k$, $k_j = n$ and $k_0 < k_1 < \cdots < k_j$. The demonstration of the previous equation can be easily verified by induction. Observe that for $j = 1$ we obtain the running time for the conventional microaggregation algorithm. As a consequence, the relative running time is defined as follows

$$\tau_j = \frac{t_j}{t_1} = \frac{k_0}{k_j} \sum_{i=1}^{j} \frac{k_i}{k_{i-1}}.$$

To optimize the normalized running time, we employ the arithmetic-mean geometric-mean inequality

$$\frac{1}{j} \sum_{i=1}^{j} a_j \geqslant \sqrt[j]{\prod_{i=1}^{j} a_j},$$

with equality if and only if all the terms a_j are equal. Thus, the optimal value of the normalized running time is given when all the terms of the summation are equal, that is,

$$\frac{k_i^\star}{k_{i-1}^\star} = E.$$

Notice that we have named E as the expansion factor since it gives the relation between any two consecutive optimal values of $k_i^\star$. Using this expression recursively, we can relate all the $k_i^\star$ values with the desired k-anonymity as follows $k_i^\star = E^i k_0$.
We shall use this expression in the specific case when $i = j$ to calculate the expansion factor

$$E = \sqrt[j]{\frac{k_j}{k_0}} = \sqrt[j]{\frac{n}{k}}.$$

As seen above, there is a bijective relation between the expansion factor E and the number of iterations j to be done until the k-anonymity is achieved. For a fixed expansion factor the number of iterations is

$$j = \frac{\ln \frac{n}{k}}{\ln E} = \log_E \frac{n}{k}.$$

Since j has to be an integer value, we have to truncate it and then recalculate the expansion factor. Finally, using the above expressions we see that the optimal value for the relative running time can be expressed

either based on the number of iterations j or depending on the expansion factor E

$$\frac{t_j^\star}{n} = jE = j\sqrt[j]{\frac{n}{k}} = \frac{E}{\ln E}\ln\frac{n}{k} = E\log_E\frac{n}{k}.$$

If we fix the number of iterations j, we have that the running time is in the form $\Theta\left(n^{(j+1)/j}\right)$, that is a subquadratic form for any $j > 1$ (at least one prepartition). On the other hand, if we fix the expansion factor E, the running time is in the form $\Theta(n\log n)$ which is a quasilinear form. The reason behind the term quasilinear is that $\log^r n = o(n^\epsilon)$ for any real values $r \geqslant 0$ and $\epsilon > 0$ when n approaches to infinity, so quasilinear denotes faster running time than any exponent strictly greater than 1, which would yield linearity. Notice that in both cases the running time is faster than the conventional algorithm which is in the form $\Theta(n^2)$. Therefore, fixing the value of the expansion factor E seems to be the best option for big datasets in terms of the running time, but it is not clear what price to pay in terms of distortion. It is beyond the scope of this work to perform a mathematical analysis on the relationship between the distortion and the parameters E and j. In the next section, we discuss the relationship between these two parameters and the distortion, based on experimental results.

Compared with the running time t_1 without prepartitioning, the relative duration of the entire process with j recursive prepartitioning stages is

$$\tau_j^\star = \frac{t_j^\star}{t_1} = j\left(\frac{k}{n}\right)^{1-\frac{1}{j}}.$$

As an example, consider $n = 10^5$ records and an anonymity parameter $k = 10$, for two-stage microaggregation, represented by $j = 2$, corresponding to a macroaggregation as prepartitioning and the microaggregation of each macrocell separately. The optimal macrocell size for prepartitioning would be $k_1^\star = \sqrt{kn} = 10^3$. This means that the dataset of $n = 10^5$ records would be first divided into $n/k_1^\star = E = 10^2$ macrocells, each of size $k_1^\star = 10^3$, and each macrocell would be microaggregated individually, with an anonymity parameter $k = 10$, resulting in $\frac{k_1^\star}{k} = E = 10^2$ microcells per macrocell. Note that the optimal solution preserves the size ratios involved.

Now, to realize the potential of this optimization, compute the relative time of the entire two-stage process, considered here without any sort of parallelization whatsoever. According to the previous mathematical analysis, the optimized relative time is a fraction $\tau_2^\star = 2\sqrt{k/n} = 2\times 10^{-2}$ of the original time. In other words, our approach with optimized prepartitioning would run 50 times faster than traditional microaggregation without prepartitioning.

In general, from the previous expressions, it is obvious that the expansion factor will not be an integer and, therefore, the $k_i^\star$ values neither. As a consequence of rounding the $k_i^\star$ values, there is a difference between the obtained running time compared to the theoretical optimal value. We can deduce the relationship between the relative error of the running time and the relative error of the used k_i values by applying $k_i = k_i^\star(1+\epsilon_i)$ which leads to

$$\epsilon_t = \frac{t_j}{t_j^\star} - 1 = \frac{1}{j}\sum_{i=1}^{j}\frac{1+\epsilon_i}{1+\epsilon_{i-1}} - 1.$$

Now, we can go further by calculating the number of iterations j that optimizes the relative running time. To do so, we derive $t_j^\star$ with respect to j, and equals it to zero to finally obtain the following

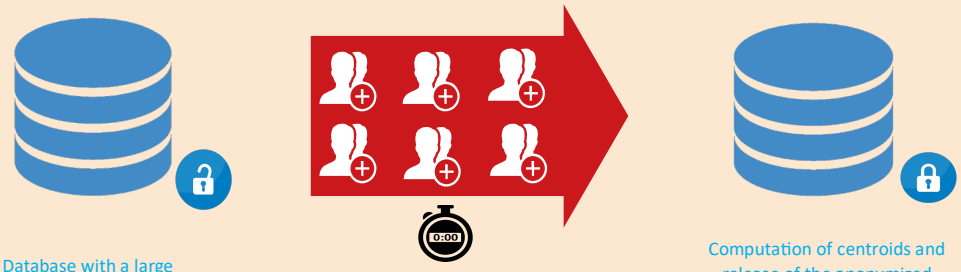
$$j^\star = \ln\frac{n}{k}.$$

In this case, the expansion factor becomes the Euler's number e and the normalized running time will be

$$t^{\star\star} = n\,e\ln\frac{n}{k}.$$

The double star notation indicates a double optimization, on the one hand the k_i values have been optimized and on the other, the number of iterations are optimized too. A summary in greater detail is offered in Fig. 6, which should serve as a convenient recapitulation of our proposal.

Brief recapitulation of our proposal to speed up the microaggregation process by means of optimized prepartitioning



Database with a large number of records n

Fast k-Anonymous microaggregation on the QIDs by applying an optimized prepartitioning strategy

Computation of centroids and release of the anonymized database

To drastically reduce the running time required to microaggregate large datasets, we present two approaches, each carrying a different price on data utility. Although the partition parameters for each choice are optimized mathematically, choosing between the two approaches has an impact on distortion, which must be assessed experimentally for a given dataset.

The first choice consists in fixing the number j of recursive partitioning stages, where $j = 1$ represents a single microaggregation process without any prepartitioning, and $j = 2$ indicates the usual approach of two aggregation stages, one of them corresponding to prepartitioning into macrocells, and the other k-anonymizing each part.

- For a number of records n up to a million, $j = 2$ or $j = 3$ should be appropriate choices with reasonable distortion impact and drastic speed-up, as our experiments confirm.
- Our mathematical analysis determines then the optimal expansion factor $E = \sqrt[j]{n/k}$ for a selected number j of stages and an anonymity parameter k.
- This gives the optimal cell size, or number of records per cluster on each iteration i, as $k_i^\star = \lfloor E^i k \rfloor$, for $i = 1, \cdots, j-1$.
- Starting with $k_{j-1}^\star$, we microaggregate recursively the dataset creating on each iteration a partition with $k_i^\star$ records per cluster. In the last recursive stage we form clusters with (at least) k records, thus satisfying the k-anonymity constraint.

The second approach consists in fixing the expansion factor E, that is, the ratio between the cardinality of a partitioning macrocell and the cardinality of its embedded microcells, for each recursive partitioning stage. An expansion factor E large enough should keep the additional distortion due to prepartitioning in check.

- Our mathematical analysis determines the corresponding number of recursive stages $j = \log_E(n/k)$, but the prepartitioning recursion is otherwise identical to the first approach.

For either approach, the optimized running time per record is estimated as $t_j^\star/n = jE = j\sqrt[j]{n/k} = E\log_E(n/k)$. The asymptotic complexity of the total running time $t_j^\star$ with the number n of records can be viewed as $\Theta\left(n^{(j+1)/j}\right)$ in terms of the number j of partitioning stages (quadratic for the conventional case $j = 1$), or as $\Theta(n \log n)$ when the expansion factor E is fixed.

Figure 6: Brief recapitulation of our proposal to speed up the microaggregation process by applying optimized prepartitioning strategies.

4. Experimental Results

In this experimental section, we aim to confirm the efficiency of our prepartitioning strategy applied for k-anonymous microaggregation, in terms of time gain and relative distortion. Keep in mind that although our novel method is illustrated with the special case of MDAV (Domingo-Ferrer & Torra (2005); Domingo-Ferrer et al. (2006); Hundepool et al. (2007); Templ (2008)), which is, one of the best-known and most widely used fixed-size microaggregation algorithms for numerical data, our novel method would be easily applied to other microaggregation algorithms. Furthermore, all experiments in their entirety were implemented and executed in Matlab 2017b, explicitly disabling any form of parallelization for fair and clear comparison.

Additionally, two datasets are considered to evaluate the performance of our novel method versus MDAV, one is standardized and the other one is synthetic. Although the computational cost will not be vary with the type of dataset processed, we are still interested in evaluating the impact of our method in terms of utility (distortion) and in comparing its performance with that of the original MDAV. In this way, a user could choose the strategy that achieves a better tradeoff between computational cost and distortion in a given context. The synthetic dataset generated randomly with a Gaussian distribution has different amount of records with 10 numerical attributes on each. The "USA House" dataset contains 5,967,303 records with 13 numerical attributes that, in this contribution, will be considered as quasi-identifiers. The tests have been performed using different subset sizes of 10^3, 10^4, 10^5 and 10^6 samples randomly selected from the aforementioned datasets. Furthermore, we stick to the common practice of normalizing each attribute of the dataset for zero-mean and unit variance.

We shall measure the relative performance gain $\tau_j \stackrel{\text{def}}{=} t'_j/t$, defined as the execution time t'_j of the novel method with j steps of prepartitioning with respect to the time t of the traditional microaggregation procedure MDAV. In this manner, relative running times τ_j will potentially range from 0% to 100%, where 100% indicates a running time identical to that of MDAV. Similarly, we shall report the incurred relative distortion $\delta_j \stackrel{\text{def}}{=} \mathcal{D}'_j/\mathcal{D}$, where $\mathcal{D}'_j$ is the distortion corresponding to the novel method with j steps of prepartitioning, and $\mathcal{D}$ is the distortion corresponding to conventional MDAV. Again, relative distortion δ_j will potentially be greater than 100%, where 100% indicates a distortion equal to conventional MDAV.

4.1. Validation of the Theoretical Model

To validate the theoretical model described in §3.2, the relative performance gain τ_1 for the case of single-stage prepartitioning method has been computed by sweeping the value of k_1, using different subset sizes of 10^3, 10^4, 10^5 and 10^6 samples randomly selected from both, the synthetic and standardized dataset.

Fig. 7 shows the value of the relative performance gain τ_1 achieved with an anonymity value $k = 10$ in each created cluster. It can be observed that the shape of the curve resembles the theoretical result, mainly for large datasets. Namely, the experimental curves get closer to the theoretical one (drawn as a dotted line) as n gets larger.

In Fig. 8 it can be observed that the minimum of τ_2 is at the expected position $k_1^\star = \sqrt{kn}$, but its value is greater than the theoretical result. The stepped form seen in Fig. 7 is due to the MDAV algorithm itself. In this algorithm, the records are grouped in clusters and when the number of records pending to be assigned to a cluster is less than $2k$, all of them are assigned to the same cluster. Therefore, if $k_1 > n/2$, only a single cluster can be made, which is equivalent to not doing any prepartition. If $n/3 < k_1 < n/2$ it is only possible to make two clusters in the prepartition and so on. This is basically the reason for the stepped behavior that appears in Fig. 7 for $k_1 = n/2, n/3, n/4, \ldots$

It can be clearly seen that the differences between both datasets almost disappear when the number of records is large enough.

4.2. Data Utility Loss

The idea behind applying an optimized prepartitioning strategy is to attain a significant reduction in the running time required by k-anonymous microaggregation, at the expense of a relatively moderate degradation in data utility. We hasten to stress that the additional degradation provoked by our method is measured in this work with respect to the distortion caused by the original MDAV.

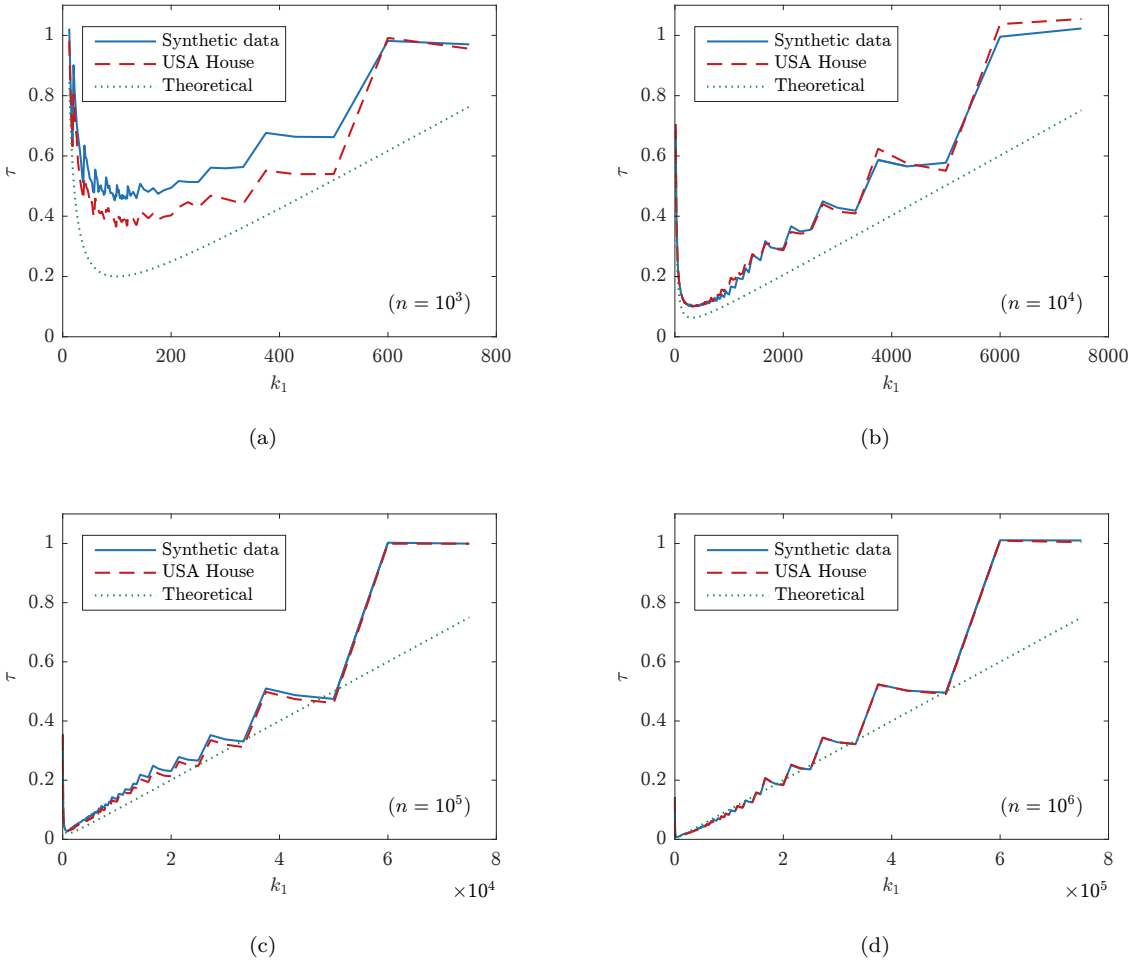
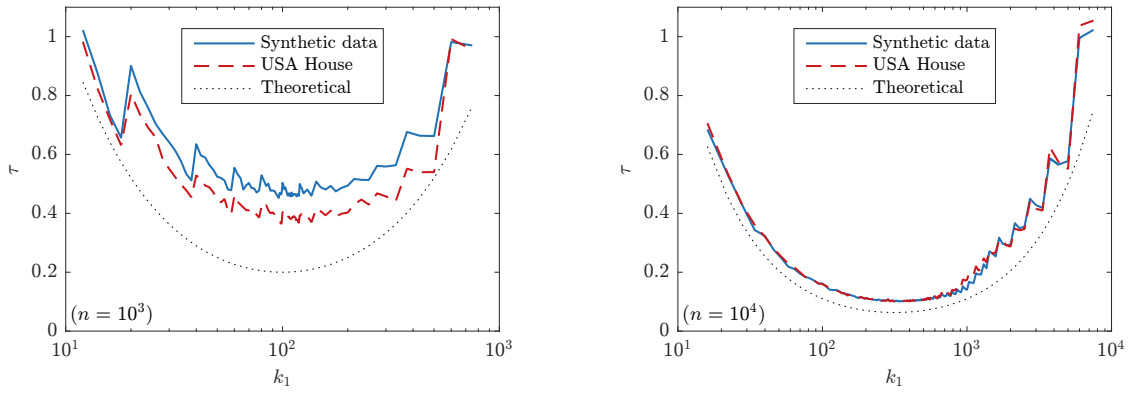
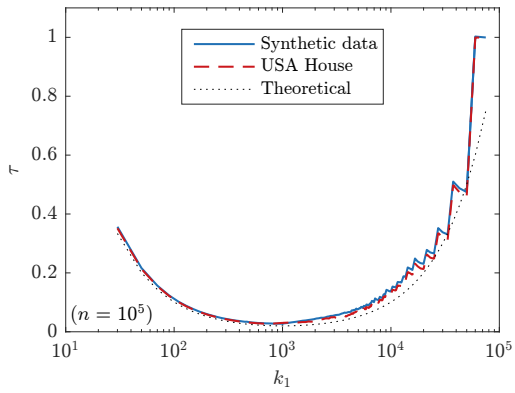


Figure 7: Relative performance gain when the value k_1 is swept, in a single-stage prepartitioning with $k = 10$, for the USA House dataset and a synthetic dataset, compared to the theoretical result, for different dataset sizes (a) 10^3 records, (b) 10^4 records, (c) 10^5 records and (d) 10^6 records.

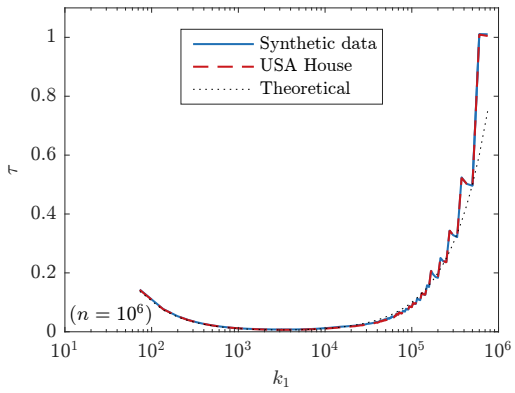


(a)

(b)



(c)



(d)

Figure 8: Relative performance gain when the value k_1 is swept, in a single-stage prepartitioning with $k = 10$, for the USA House dataset and a synthetic dataset, compared to the theoretical result, for different dataset sizes (a) 10^3 records, (b) 10^4 records, (c) 10^5 records and (d) 10^6 records. The logarithmic axis shows that the optimal k_1 value is approximately the geometric mean between k and n.

Although this distortion might not be as low as expected when optimizing prepartitioning to get minimum running times, we feel that the benefits of this significant acceleration outweigh the extra distortion on current application domains of data.

Fig. 9 shows the relative values of distortion and runtime for a different number of steps (from $j = 1$ to $j^\star$), when the optimal values of $k_j^\star$ are used.

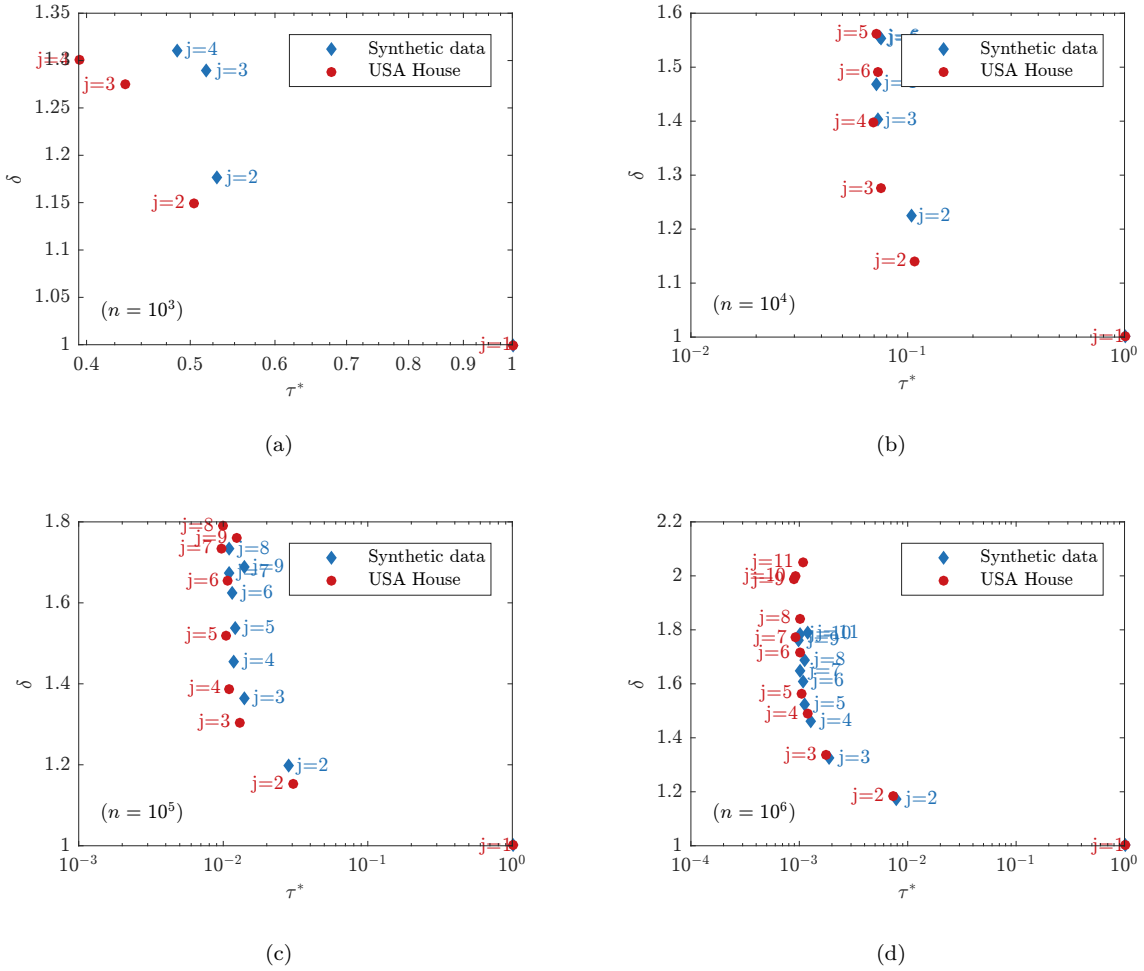


Figure 9: Relative distortion vs. the optimal relative running time, computed for both, the USA House dataset and a synthetic dataset with (a) 10^3 records, (b) 10^4 records, (c) 10^5 records and (d) 10^6 records.

It has to be noticed that from $j = 3$ towards above, there is no substantial improvement in the execution time and at the same time a significant increment in the utility loss is being obtained. As stated in the previous section, there is a bijective relation between the number of iterations j and the expansion factor E. For the optimal value of iterations $j^\star$, the expansion factor is the number $e \approx 2.7183$. In this special case when $E = e$, it is necessary for E to be truncated, thus, the relation between two consecutive $k_i^\star$ values will be $k_i^\star = 2k_{i-1}^\star$, that is, in each step, the cluster will be split into half. Even if this optimal case considerably reduces the execution time required by k-anonymous microaggregation, it has a very negative effect on the information utility.

In order to illustrate this negative effect, we have computed δ_j for different expansion factors, as shown in Fig. 10.

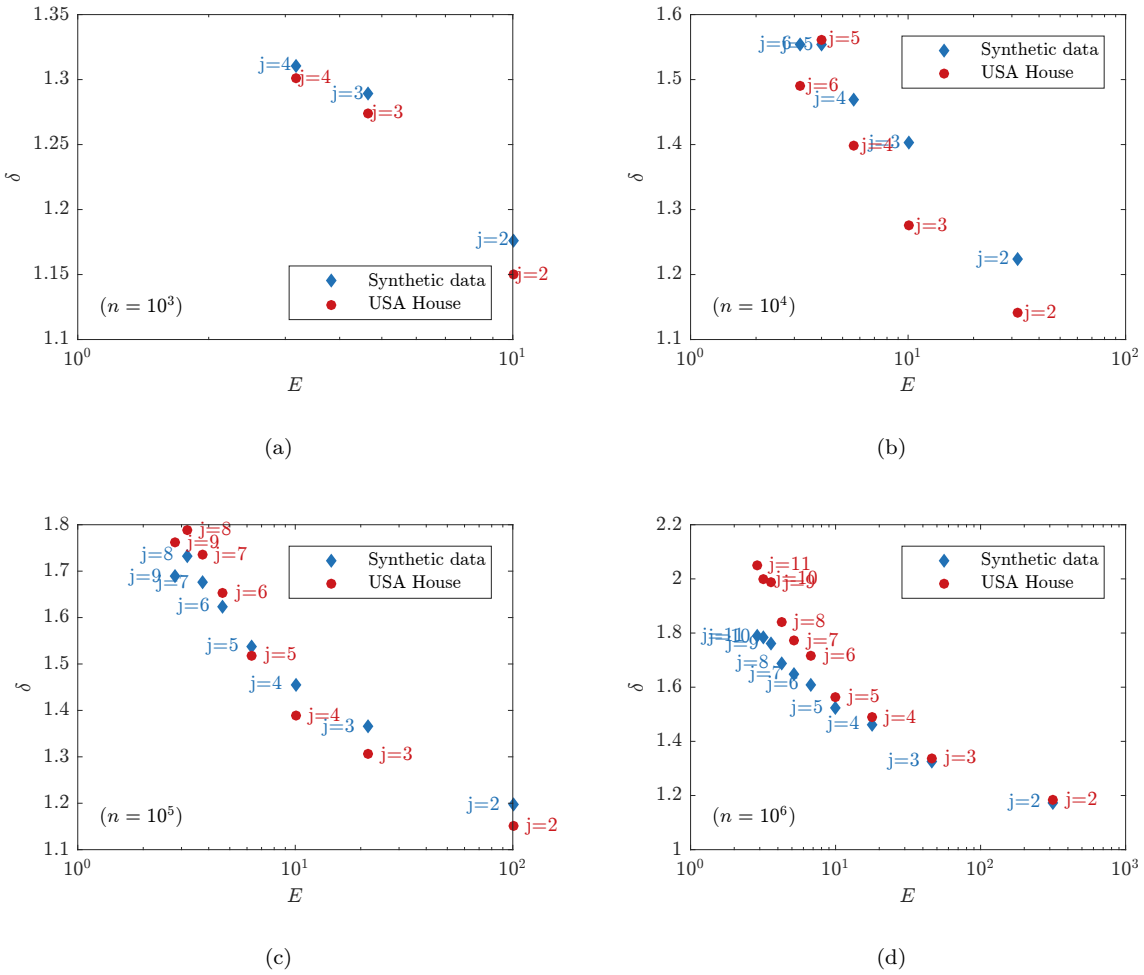


Figure 10: Relative distortion vs. expansion factor, when the running time is optimized, computed for both, the USA House dataset and a synthetic dataset, with (a) 10^3 records, (b) 10^4 records, (c) 10^5 records and (d) 10^6 records.

From Fig. 9 and Fig. 10, it can be seen that the expansion factor E and the number of iterations j influence in the relative distortion δ. A mathematical model in order to analyze the relation between these parameters is out of the scope of this work. The experimental results show that, when the expansion factor E decrease (that is, the number of iterations j increases), then the relative distortion δ increases. Additionally, Fig. 10 also suggests that the number of iterations could be a good parameter in order to predict the relative distortion. For example, in the case $j = 2$, the relative distortion δ is around the value 1.2 for $n = 10^4$, $n = 10^5$ and $n = 10^6$. Nevertheless, as mentioned above, we do not have any mathematical analysis to corroborate this claim. The case $n = 10^3$ does not allow us to draw conclusions, we have simply calculated it for completeness.

4.3. Tradeoff between Information Utility and Running Time

In the previous section, we have seen that, for values of j greater than three, there is almost no significant improvement in the optimal running time and at the same time a significant increment in the utility loss is being obtained. However, in order to find a tradeoff between the relative running time and the utility loss of the data, we have computed the relative values of distortion and running time for $j = 2$ and $j = 3$ as shown in Fig. 11 and Fig. 12. To do this, we have swept the values of k_i (k_1 for $j = 2$ and $k_1,\ k_2$ for $j = 3$).

In the case $j = 2$, we obtain a curve that starts at the point $(1, 1)$ for $k_1 = k$, which is equivalent to not performing any prepartition. Then, when k_1 grows, the running time decreases and the distortion increases until it reaches a maximum value for the distortion. We have realized that, in all our experiments for $j = 2$, the maximum distortion is reached before than the minimum running time. After this point, both, the distortion and the running time decrease until $k_1 = k_1^\star$ where we have the minimum for the running time. In the last part of the curve, that is for $k_1 > k_1^\star$, the running time increases and the distortion decreases until the starting point $(1, 1)$ is reached again when $k_1 = n$. This last section of the curve, its lower convex part, is the section where we shall find the tradeoff between the running time and the distortion. Therefore, we can state that the values of k_1 that we are interested in, are in the range $k_1^\star \leqslant k_1 < n$.

For the case $j = 3$, once k_2 has been set, the swept of k_1 draws a curve in the same way that those of the case $j = 2$. The whole set of curves when k_2 is swept, draws a cloud as the red ones shown in Fig. 11 and Fig. 12. In this case, the points we are interested in, are in the lower convex envelope of the cloud. As a consequence of the result for the case $j = 2$, we have depicted in a different color the points (τ, δ) corresponding to $k_1 \geqslant k_1^\star$ and $k_2 \geqslant k_2^\star$ for the case $j = 3$ as shown in Fig. 13. It can be seen that these points correspond to the lower convex envelope of the cloud, therefore, to improve the utility of the data, k_i values greater than the optimal ones should be used, worsening in this case the running time. The rest of the values ($k_i < k_i^\star,\ \ i = 1, 2$) are not interesting, since they worsen both the execution time and the utility of the data. Therefore, we can conclude that the more we move away from the optimum, the longer the execution time, but the less distorted the data will be.

The absolute running time of the traditional algorithm and our novel method, used in this work will certainly vary depending on both n and k, as well as on the computer and the number of cores employed. However, most of our experiments are in terms of running times relative to the traditional one MDAV.

The experiments summarized in tables 3 and 4, using our method, have been designed to guarantee a resulting distortion that is about 10% higher than that of MDAV (presented in tables 1 and 2). As shown in Table 2, the distortion when using the original MDAV over the USA House dataset is 0.0236, while the corresponding distortion is 0.0259 (Table 4) when using our method. In this example, the absolute increasing of degradation due to our method is only 0.0023, i.e., a 9.74% of the distortion obtained with MDAV. In addition, the corresponding tests are performed considering $k = 10$, which is already high since k is commonly small for SDC applications. Thus, even lower relative degradation could be expected for lower (but still practical) values of k.

As suggested by the results of the aforementioned experiments, the increase in distortion levels due to our method could take values as low as 10%. We feel that this level of relative degradation might be reasonably acceptable since it is low, particularly if the original distortion is already low. But a dramatically speeding up is obtained in return. To illustrate this statement, imagine, for instance, having a distortion of 0.11 instead of 0.10 in exchange for a speedup factor of 56 when a big data application is involved. Such

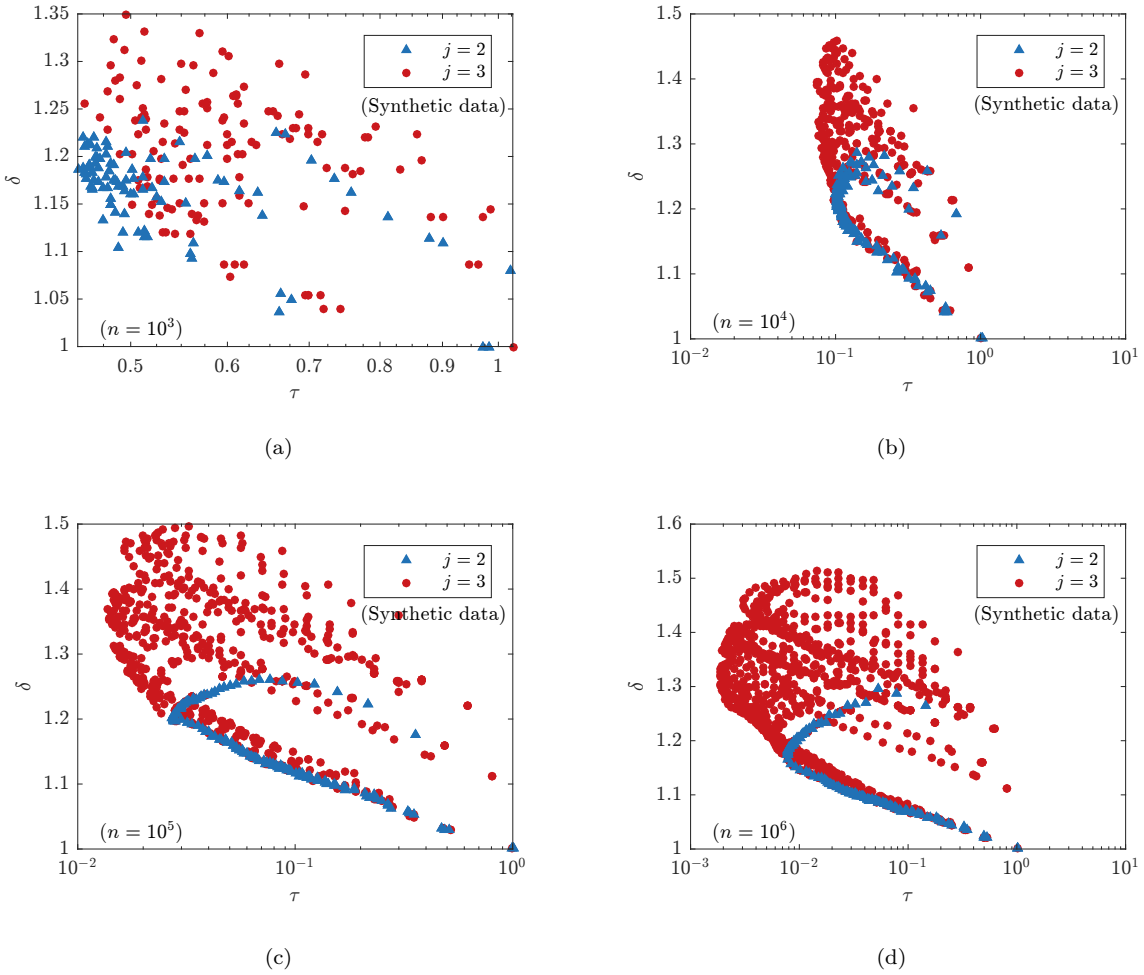


Figure 11: Relative running time vs. relative distortion for the case $j = 2$ and $j = 3$ using the synthetic dataset with sizes (a) 10^3 records, (b) 10^4 records, (c) 10^5 records and (d) 10^6 records.

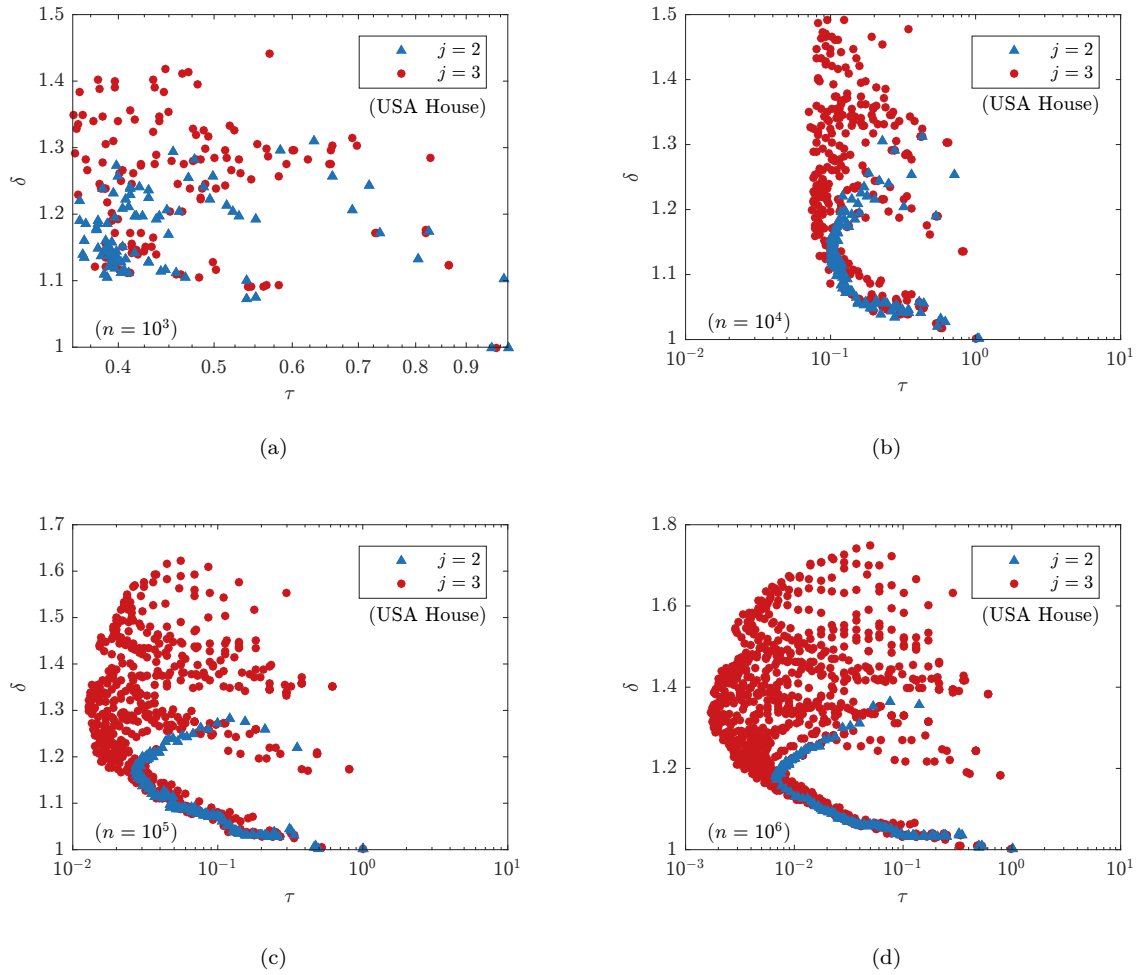


Figure 12: Relative running time vs. relative distortion for the case $j = 2$ and $j = 3$ using the USA House dataset with sizes (a) 10^3 records, (b) 10^4 records, (c) 10^5 records and (d) 10^6 records.

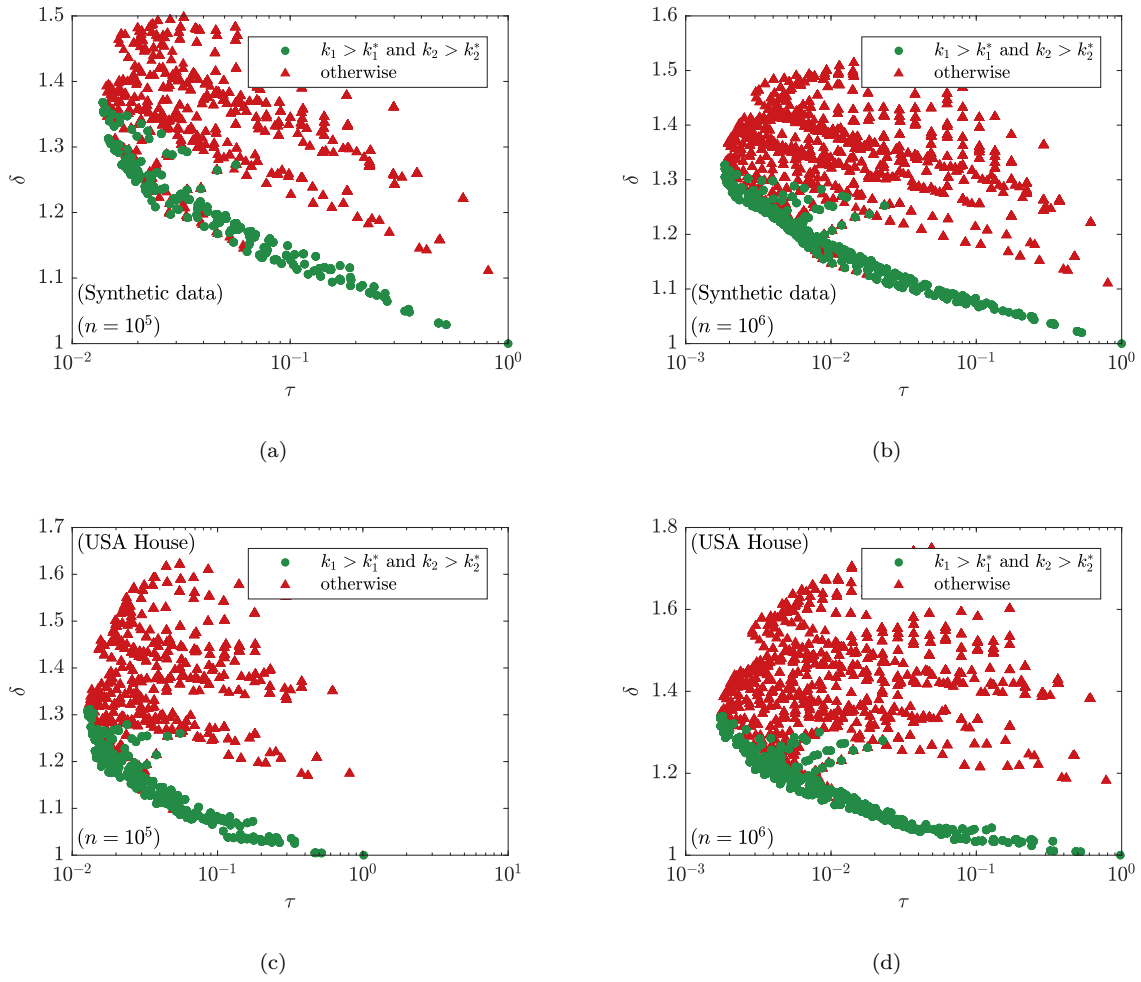


Figure 13: Relative running time vs. relative distortion for the case $j = 3$ differentiating the points for which to $k_1 \geqslant k_1^\star$ and $k_2 \geqslant k_2^\star$ from the others. These points define the lower convex envelope of the cloud. For a synthetic dataset with (a) 10^5 records and (b) 10^6 records; and for the USA House dataset with (c) 10^5 records and (d) 10^6 records.

acceleration may be very useful for big data and real-time applications, whose predominance grows year by year due to the increasing availability of information.

Undoubtedly, minimum execution times are reached in exchange for higher values of distortion. However, such a reduction in time will be required only when dealing with very large databases. Interestingly, within vast amounts of data, it is possible to obtain reasonable levels of anonymity (large enough values for k) for low distortion, since the more data there is, the more likely it is to find records with similar values for quasi-identifiers. In this line, a fraction of a distortion that is already low (say 10%) will still be low, thus the resulting data utility will be preserved (Rodríguez-Hoyos et al. (2018)).

In any case, if less distortion were required, the parameter k_1 of our method can always be tuned to satisfy the requirement of any application domain or that of the entities involved in exploiting and protecting such data.

Table 1: Reference running times and distortions without prepartitioning

Synthetic dataset	**Dataset size**			
	$n = 10^3$	$n = 10^4$	$n = 10^5$	$n = 10^6$
Running time	0.016 s	1.284 s	90.977 s	1.094×10^4 s (3hr 2min 20sec)
Distortion	0.4325	0.2788	0.1801	0.1165

Reference running times and reference distortions for MDAV without prepartitioning ($j = 1$), applied to the synthetic dataset with different number of records and $k = 10$.

Table 2: Reference running times and distortions without prepartitioning

USA House dataset	**Dataset size**			
	$n = 10^3$	$n = 10^4$	$n = 10^5$	$n = 10^6$
Running time	0.022 s	1.370 s	108.575 s	1.303×10^4 s (3hr 37min 10sec)
Distortion	0.2386	0.1212	0.0558	0.0236

Reference running times and reference distortions for MDAV without prepartitioning ($j = 1$), and our novel method, applied to the standardized dataset, USA House, with different number of records and $k = 10$.

Table 3: Absolute running times and distortions with prepartitioning

Synthetic dataset	**Dataset size and prepartition size**			
	$n = 10^3$, $k_1 = 273$	$n = 10^4$, $k_1 = 1765$	$n = 10^5$, $k_1 = 11111$	$n = 10^6$, $k_1 = 39963$
Running time	0.0089 s	0.3812 s	13.691 s	355.4178 s (5min 55.42sec)
Distortion	0.4724	0.3068	0.1982	0.1284

Absolute running times and distortions using the proposed algorithm with a single prepartition (hence $j = 2$ aggregation stages) and $k_1 > k_1^\star$ to achieve a k-anonymization of $k = 10$, for the synthetic dataset with 10^3, 10^4, 10^5 and 10^6 records. The k_1 values has been chosen to increase the distortion only about a 10% respect to MDAV.

5. Conclusion

We have presented an optimized prepartitioning method that reduces drastically the running time for the k-anonymous microaggregation algorithm. The method is based on prepartitioning a dataset recursively until the desired k-anonymity is achieved. Under the assumption that the running time of the conventional

Table 4: Absolute running times and distortions with prepartitioning

USA House dataset	**Dataset size and prepartition size**			
	$n = 10^3$, $k_1 = 300$	$n = 10^4$, $k_1 = 428$	$n = 10^5$, $k_1 = 2054$	$n = 10^6$, $k_1 = 19784$
Running time	0.01 s	0.1419 s	3.9764 s	231.731 s (3min 51.73sec)
Distortion	0.2653	0.1374	0.062	0.0259

k-anonymization algorithm scales with $t = n^2/k$, we have calculated the parameters that optimize the running time. These parameters are, the minimum number k_i of records in a cluster in each prepartition, and the number j of prepartitions to be carried out recursively until the dataset is k-anonymized. The main advantage concerning applicability of this paper is that using the proposed algorithm the running time is in the subquadratric form $\Theta(n^{(j+1)/j})$ if we fix the number of iterations j, and in the quasilinear form $\Theta(n \log n)$ if we fix the expansion factor. Both cases are faster than the conventional algorithm which is in the form $\Theta(n^2)$.

The proposed method has been implemented in Matlab 2017b using MDAV as a conventional microaggregation algorithm. Several experiments have been done in order to check the correctness of the analysis. We have used the USA House dataset and a synthetic dataset and we have found that the running time can be optimized using the calculated parameters $k_i^\star$ and $j^\star$. The main drawback of the proposal is the reduction of the utility of the data in terms of distortion. Distortion depends on the number of iterations and the expansion factor. Increasing the number of iterations, that is, reducing the expansion factor, entails an increase of the distortion. Experimental results suggest that setting the number of iterations fixes the relative distortion independently of the number of records n. In this case, we can control the distortion, but with subquadratic running time, instead of the quasilinear running time that we have if we set the expansion factor. We have obtained empirically that for more than two prepartitions, that is for $j > 3$, there is no worthwhile improvement in the running time, but there is a considerable deterioration in the utility of the data. For the cases $j = 2$ and $j = 3$, the tradeoff between the running time and the distortion is found for values of k_i greater or equal than the optimal $k_i^\star$.

The assessment of our proposal is limited in the sense that it is implemented only on top of MDAV. Its impact on data utility, if other microaggregation approaches are used, remains unknown. Thus, in future investigation, we would like to study the application of our optimized prepartitioning method to algorithms capable of outperforming MDAV. Due to the simplicity, efficiency, and more than reasonable performance of MDAV, however, our study should be at least as useful as MDAV is convenient for a variety of datasets, and the framework developed here should keep its mathematical appeal for any numerical algorithms of similar quadratic complexity $\Theta(n^2/k)$ in the number n of records and inverse in the microcell size k.

Acknowledgment

We gratefully acknowledge the invaluable assistance of Irene Carrión-Barberà, M.D., in the preparation of the medical example in Fig. 1. We would also like to thank the anonymous reviewers for their helpful suggestions to improve the readability and contents of this paper.

This manuscript presents some of the results developed through the collaboration of the Universitat Politècnica de Catalunya (UPC) and Scytl Secure Electronic Voting S.A. (Scytl) in the context of the project "Data-Distortion Framework", and in accordance with the guidelines therein. This work is thus partly supported by the Spanish Ministry of Industry, Energy and Tourism (MINETUR) through the "Acción Estratégica Economía y Sociedad Digital (AEESD)" funding plan, through the aforementioned project, "Data-Distortion Framework (DDF)", ref. TSI-100202-2013-23.

Additional funding supporting this work has been granted to UPC by the Spanish Ministry of Economy and Competitiveness (MINECO) through the "Anonymized Demographic Surveys (ADS)" project, ref. TIN2014-58259-JIN, under the funding program "Proyectos de I+D+i para Jóvenes Investigadores", and

through the project "MAGOS", ref. TEC2017-84197-C4-3-R, as well as by the Government of Catalonia, under grant 2014 SGR 1504.

References

Abidi, B., & Yahia, S. B. (2017). Generating k-anonymous microdata by fuzzy possibilistic clustering. In *Proc. Int. Conf. Database, Expert Syst. Appl. (DEXA)* (pp. 3–17). Lyon, France.

Aloise, D., & Araújo, A. (2015). A derivative-free algorithm for refining numerical microaggregation solutions. *Int. Trans. Oper. Res.*, *22*, 693–712.

AOL (2006). AOL search data scandal. URL: http://en.wikipedia.org/wiki/AOL_search_data_scandal.

Arres, B., Kabachi, N., & Boussaid, O. (2015). A data pre-partitioning and distribution optimization approach for distributed data warehouses. In *Proc. Int. Conf. Parallel, Distrib. Process. Tech., Appl. (PDPTA)* (pp. 454–461). Las Vegas, NV.

Brickell, J., & Shmatikov, V. (2008). The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In *Proc. ACM SIGKDD Int. Conf. Knowl. Discov., Data Min. (KDD)*. Las Vegas, NV.

Cukier, K. (2014). Big data is better data. Technol., Entertain., Design (TED) Talk. URL: www.ted.com/talks/kenneth_cukier_big_data_is_better_data.

Dankar, F. K., Brien, R., Adams, C., & Matwin, S. (2014). Secure multi-party linear regression. In *Proc. Int. Jnt. Conf. Ext. Database Technol., Database Theory (EDBT/ICDT)* (pp. 406–414). Athens, Greece.

Defays, D., & Nanopoulos, P. (1993). Panels of enterprises and confidentiality: The small aggregates method. In *Proc. Symp. Design, Anal. Longit. Surv., Stat. Can.* (pp. 195–204). Ottawa, Canada.

Domingo-Ferrer, J., & González-Nicolás, Ú. (2010). Hybrid microdata using microaggregation. *Inform. Sci.*, *180*, 2834–2844.

Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J. M., & Sebé, F. (2006). Efficient multivariate data-oriented microaggregation. *VLDB J.*, *15*, 355–369.

Domingo-Ferrer, J., & Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl., Data Eng.*, *14*, 189–201.

Domingo-Ferrer, J., Sebé, F., & Solanas, A. (2008). A polynomial-time approximation to optimal multivariate microaggregation. *Comput., Math., Appl.*, *55*, 714–732.

Domingo-Ferrer, J., Solanas, A., & Castellà-Roca, J. (2009). $h(k)$-private information retrieval from privacy-uncooperative queryable databases. *Online Inform. Rev.*, *33*, 720–744.

Domingo-Ferrer, J., & Torra, V. (2003). Fuzzy microaggregation for microdata protection. *J. Adv. Comput. Intell., Intell. Inform.*, *7*.

Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Min., Knowl. Discov.*, *11*, 195–212.

Domingo-Ferrer, J., & Torra, V. (2008). A critique of k-anonymity and some of its enhancements. In *Proc. Workshop Priv., Secur., Artif. Intell. (PSAI)* (pp. 990–993). Barcelona, Spain.

Dwork, C. (2006). Differential privacy. In *Proc. Int. Colloq. Automata, Lang., Program. (ICALP)* (pp. 1–12). Venice, Italy volume 4052 of *Lect. Notes Comput. Sci. (LNCS)*.

Fayyoumi, E., & Nofal, O. (2018). Applying genetic algorithms on multi-level micro-aggregation techniques for secure statistical databases. In *Proc. ACS/IEEE Int. Conf. Comput. Syst., Appl. (AICCSA)* (pp. 1–6). Aqaba, Jordan.

Gursoy, M. E., Inan, A., Nergiz, M. E., & Saygin, Y. (2017). Privacy-preserving learning analytics: Challenges and techniques. *IEEE Trans. Learn. Technol.*, *10*, 68–81.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE J. Intell. Syst.*, *24*, 8–12.

Hundepool, A., de Wetering, A. V., Ramaswamy, R., Franconi, L., Capobianchi, A., de Wolf, P.-P., Domingo-Ferrer, J., Torra, V., Brand, R., & Giessing, S. (2003). *μ-ARGUS version 3.2 software and user's manual*. Stat. Neth. Voorburg, Netherlands. URL: http://neon.vb.cbs.nl/casc.

Hundepool, A., Ramaswamy, R., DeWolf, P.-P., Franconi, L., Brand, R., & Domingo-Ferrer, J. (2007). *μ-ARGUS version 4.1 software and user's manual*. Stat. Neth. Voorburg, Netherlands. URL: http://neon.vb.cbs.nl/casc.

Iftikhar, M., Wang, Q., & Lin, Y. (2019). Publishing differentially private datasets via stable microaggregation. In *Proc. Int. Jnt. Conf. Ext. Database Technol., Database Theory (EDBT/ICDT)* (pp. 662–665). Lisbon, Portugal.

Inan, A., Kantarcioglu, M., & Bertino, E. (2009). Using anonymized data for classification. In *Proc. IEEE Int. Conf. Data Eng. (ICDE)* (pp. 429–440). Shanghai, China.

Ke, Q., Prabhakaran, V., Xie, Y., Yu, Y., Wu, J., & Yang, J. (2011). Optimizing data partitioning for data-parallel computing. In *Proc. USENIX Workshop Hot Topics Oper. Syst. (HotOS)*. Napa, CA.

Laszlo, M., & Mukherjee, S. (2005). Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Trans. Knowl., Data Eng.*, *17*, 902–911.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. In *Proc. IEEE Int. Conf. Data Eng. (ICDE)* (pp. 25–35). Atlanta, GA.

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy beyond k-anonymity and l-diversity. In *Proc. IEEE Int. Conf. Data Eng. (ICDE)* (pp. 106–115). Istanbul, Turkey.

Lin, J. L., Wen, T. H., Hsieh, J. C., & Chang, P. C. (2010). Density-based microaggregation for statistical disclosure control. *Expert Syst., Appl.*, *37*, 3256–3263.

Liu, H., Zhang, Q., Guo, K., & Wu, Y. (2018). Grey maximum distance to average vector based on quasi-identifier attribute. *J. Grey Syst.*, *30*.

Mahmood, A. N., Kabir, M. E., & Mustafa, A. K. (2012). New multi-dimensional sorting based k-anonymity microaggregation for statistical disclosure control. In *Proc. EAI Int. Conf. Secur., Priv. Commun. Netw. (SecureComm)* (pp. 256–272). Padua, Italy.

Matatov, N., Rokach, L., & Maimon, O. (2010). Privacy-preserving data mining: A feature set partitioning approach. *Inform. Sci.*, *180*, 2696–2720.

Matwin, S., Nin, J., Sehatkar, M., & Szapiro, T. (2015). A review of attribute disclosure control. In G. Navarro-Arribas, & V. Torra (Eds.), *Advanced research in data privacy* (pp. 41–61). Switzerland: Springer Int. Publ. volume 567 of *Stud. Comput. Intell.*.

Mohamad Mezher, A., García-Álvarez, A., Rebollo-Monedero, D., & Forné, J. (2017). Computational improvements in parallelized k-anonymous microaggregation of large databases. In *Proc. IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS), Workshop Priv., Secur. Big Data (PSBD)* (pp. 258–264). Atlanta, GA.

Mortazavi, R., & Jalili, S. (2017). Fine granular proximity breach prevention during numerical data anonymization. *Trans. Data Priv.*, *10*, 117–144.

Mortazavi, R., Jalili, S., & Gohargazi, H. (2014). Fast data-oriented microaggregation algorithm for large numerical datasets. *Knowl.-Based Syst.*, *67*, 195–.205.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proc. IEEE Symp. Secur., Priv. (S&P)* (pp. 111–125). Oakland, CA.

Parra-Arnau, J., Domingo-Ferrer, J., & Soria-Comas, J. (2019). Differentially private data publishing via cross-moment microaggregation. *Inform. Fusion*, . In press.

Rebollo-Monedero, D., Forné, J., & Domingo-Ferrer, J. (2008). From t-closeness to PRAM and noise addition via information theory. In *Proc. Int. Conf. Priv. Stat. Databases (PSD)* Lect. Notes Comput. Sci. (LNCS) (pp. 100–112). Istanbul, Turkey.

Rebollo-Monedero, D., Forné, J., & Domingo-Ferrer, J. (2010). From t-closeness-like privacy to postrandomization via information theory. *IEEE Trans. Knowl., Data Eng.*, *22*, 1623–1636. URL: http://doi.org/10.1109/TKDE.2009.190.

Rebollo-Monedero, D., Forné, J., Pallarès, E., & Parra-Arnau, J. (2013a). A modification of the Lloyd algorithm for k-anonymous quantization. *Inform. Sci.*, *222*, 185–202. URL: http://doi.org/10.1016/j.ins.2012.08.022.

Rebollo-Monedero, D., Forné, J., & Soriano, M. (2011). An algorithm for k-anonymous microaggregation and clustering inspired by the design of distortion-optimized quantizers. *Data, Knowl. Eng.*, *70*, 892–921. URL: http://doi.org/10.1016/j.datak.2011.06.005.

Rebollo-Monedero, D., Forné, J., Soriano, M., & Puiggalí Allepuz, J. (2017). p-Probabilistic k-anonymous microaggregation for the anonymization of surveys with uncertain participation. *Inform. Sci.*, *382–383*, 388–414. URL: http://doi.org/10.1016/j.ins.2016.12.002.

Rebollo-Monedero, D., Hernández-Baigorri, C., Forné, J., & Soriano, M. (2018). Incremental k-anonymous microaggregation in large-scale electronic surveys with optimized scheduling. *IEEE Access*, *6*, 60016–60044. URL: http://doi.org/10.1109/ACCESS.2018.2875949.

Rebollo-Monedero, D., Mohamad Mezher, A., Casanova Colomé, X., Forné, J., & Soriano, M. (2019). Efficient k-anonymous microaggregation of multivariate numerical data via principal component analysis. *Inform. Sci.*, *503*, 417–443. URL: http://doi.org/10.1016/j.ins.2019.07.042. In press.

Rebollo-Monedero, D., Parra-Arnau, J., Díaz, C., & Forné, J. (2013b). On the measurement of privacy as an attacker's estimation error. *Int. J. Inform. Secur.*, *12*, 129–149. URL: http://doi.org/10.1007/s10207-012-0182-5.

Rodríguez-Hoyos, A., Estrada-Jiménez, J., Rebollo-Monedero, D., Parra-Arnau, J., & Forné, J. (2018). Does k-anonymous microaggregation affect machine-learned macrotrends? *IEEE Access*, *6*, 28258–28277. URL: http://doi.org/10.1109/ACCESS.2018.2834858.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *Amer. Psychol.*, *44*, 1276–1284.

Salas, J., & Torra, V. (2018). A general algorithm for k-anonymity on dynamic databases. In J. García-Alfaro, J. Herrera-Joancomartí, G. Livraga, & R. Ríos (Eds.), *Data privacy management, cryptocurrencies and blockchain technology* (pp. 407–414). Switzerland: Springer volume 11025 of *Lect. Notes Comput. Sci. (LNCS)*.

Samarati, P. (2001). Protecting respondents' identities in microdata release. *IEEE Trans. Knowl., Data Eng.*, *13*, 1010–1027.

Sánchez, D., Domingo-Ferrer, J., Martínez, S., & Soria-Comas, J. (2016). Utility-preserving differentially private data releases via individual ranking microaggregation, . *30*, 1–14.

Sankar, L., Rajagopalan, S. R., & Poor, H. V. (2013). Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Trans. Inform. Forensics, Secur.*, *8*, 838–852.

Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., , & Martínez, S. (2014). Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *VLDB J.*, *23*, 771–794.

Sun, X., Wang, H., Li, J., & Truta, T. M. (2008). Enhanced p-sensitive k-anonymity models for privacy preserving data publishing. *Trans. Data Priv.*, *1*, 53–66.

Sun, X., Wang, H., Li, J., & Zhang, Y. (2012). An approximate microaggregation approach for microdata protection. *Expert Syst., Appl.*, *39*, 2211–2219.

Sweeney, L. (2000a). *Simple demographics often identify people uniquely*. Work. Paper 3 Carnegie Mellon Univ.

Sweeney, L. (2000b). *Uniqueness of simple demographics in the U.S. population*. Tech. Rep. LIDAP-WP4 Carnegie Mellon Univ., Sch. Comput. Sci., Data Priv. Lab. Pittsburgh, PA.

Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. *Int. J. Uncertain., Fuzz., Knowl.-Based Syst.*, *10*, 557–570.

Tabik, S., Ortega, G., Garzón, E. M., & Suárez, D. (2016). A data partitioning model for highly heterogeneous systems. In *Proc. Int. Eur. Conf. Parallel, Distrib. Comput. (Euro-Par)* (pp. 468–479). Grenoble, France.

Templ, M. (2008). Statistical disclosure control for microdata using the R-package sdcMicro. *Trans. Data Priv.*, *1*, 67–85. URL: http://cran.r-project.org/web/packages/sdcMicro.

Templ, M. (2017). *Statistical disclosure control for microdata: Methods and applications in R*. Cham, Switzerland: Springer Int. Publ.

Templ, M., Meindl, B., Kowarik, A., & Chen, S. (2014). *Introduction to statistical disclosure control (SDC)*. Work. Paper 7 Int. Househ. Surv. Netw. (IHSN). URL: www.ihsn.org/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf.

Torra, V., & Navarro-Arribas, G. (2018). Probabilistic metric spaces for privacy by design machine learning algorithms: Modeling database changes. In *Data privacy management, cryptocurrencies and blockchain technology* (pp. 422–430). Switzerland: Springer volume 11025 of *Lect. Notes Comput. Sci. (LNCS)*.

Truta, T. M., & Vinay, B. (2006). Privacy protection: p-Sensitive k-anonymity property. In *Proc. Int. Workshop Priv. Data Mgmt. (PDM)* (p. 94). Atlanta, GA.

Vaidya, J., Clifton, C. W., & Zhu, Y. M. (2006). *Privacy preserving data mining*. New York, NY: Springer.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv Prepr.*, . URL: http://arxiv.org/abs/1606.05718.

Wei, R., Tian, H., & Shen, H. (2018). Improving k-anonymity based privacy preservation for collaborative filtering. *Comput., Elect. Eng.*, *67*, 509–519.

Zhang, Z., Wang, X., Uden, L., Zhang, P., & Zhao, Y. (2018). e-DMDAV: A new privacy preserving algorithm for wearable enterprise information systems. *Enterp. Inform. Syst.*, *12*, 492–504.